MÁSTER EN ESTADÍSTICA APLICADA PARA LA CIENCIA DE DATOS CON R SOFTWARE

Trabajo Fin de Máster

Métodos estadísticos aplicados a indicadores socioeconómicos de la República Argentina

By

Javier Alejandro Di Salvo

DIRECTOR: Juan Luis López Garrancho

FECHA: 01/10/2023

ENTIDAD COLABORADORA: Máxima Formación S.L.

RESUMEN

En este trabajo de investigación se pretende evaluar cuales son las causas que afectan el Producto Bruto Interno en las diferentes provincias de la República Argentina, en relación a un conjunto de indicadores socioeconómicos.

Para lograr este objetivo se implementan distintas herramientas estadísticas como por ejemplo: pruebas paramétricas, modelos de regresión avanzados, métodos de "clusters" y análisis de componentes principales. Este estudio permite conocer que magnitudes son influyentes para el PBI como así también que regiones presentan características socioeconómicas similares.

Palabras clave: pruebas paramétricas, modelos de regresión lineal múltiple, "clusters" y análisis de componentes principales.

AGRADECIMINETOS

A Juan Luis López por su dedicación y acompañamiento para realizar este trabajo.

Contents

INTRODUCCIÓN

En este trabajo se indaga acerca del efecto del PBI per cápita en relación a diversos indicadores tanto sociales como económicos en las distintas provincias de la República Argentina durante el período 2018. El PBI per cápita es uno de las métricas más importantes al momento de evaluar el nivel de desarrollo de una economía por habitante, dado que permite identificar que regiones necesitan mayor apoyo en materia de políticas públicas y cuales presentan una situación económica más favorable. En particular, los países con altos índices de PBI p/c son considerados desarrollados, mientras que aquellos con niveles bajos de PBI p/c son llamados subdesarrollados. Las economías emergentes, cuyo nivel de producción de bienes y servicios es intermedio son considerados países en vías de desarrollo.

La relación existente entre la desigualdad social y el desarrollo económico regional, ya ha sido abordada en otras investigaciones donde se advierte cuáles son las regiones más desfavorecidas por ejemplo, en materia de: infraestructura, pobreza, acceso a la educación y al sistema de salud [1], [2], [3].

Esta temática es tratada con el objetivo de profundizar sobre el efecto del PBI per cápita y el nivel socioeconómico de las diferentes provincias de la República Argentina y determinar si existen provincias con comportamiento similar e interpretar sus causas. Para ello, el estudio se lleva a cabo en cinco etapas. En la primera fase, se efectúa un análisis exploratorio de la base de datos con el fin de recopilar información estadística básica. Luego, en la segunda etapa, se implementan distintas técnicas estadísticas para evaluar, que regiones y provincias comparten indicadores socioeconómicos e inferir cuáles son sus causas. En la tercera fase se estudia el comportamiento del PBI per cápita en relación a un conjunto de predictores. En la cuarta etapa se evalúan conglomerados de provincias que comparten indicadores socioeconómicos. En la quinta fase, se analiza la relación entre provincias y variables mediante un análisis de componentes principales. Por último, se obtienen resultados preliminares y se formulan conclusiones.

Este ensayo contribuirá a poder identificar cuáles son los indicadores socioeconómicos que inciden en las regiones del país más afectadas en relación a su nivel de desarrollo económico, lo cual podría ayudar a los gobiernos locales a implementar políticas públicas de desarrollo regional mejorando la calidad de vida de la población.

MATERIAL y MÉTODOS

El material utilizado para realizar este análisis, parte de una base de datos construida por el Banco Mundial, donde se recolectan los principales indicadores socioeconómicos de 22 provincias de la República Argentina en el 2018.

La metodología de trabajo está estructurada en las siguientes etapas:

- selección de variables,
- análisis preliminar de los datos,
- creación de un modelo de regresión lineal múltiple y logística,
- agrupación de provincias mediante "clusters",
- reducción de la dimensionalidad del conjunto de magnitudes y de provincias utilizando un análisis de componentes principales.

Las características de las magnitudes analizadas se muestran en el siguiente cuadro:

Magnitud	Descripción	Tipo de variable	Unidad de medida
province	Provincia.	Alfabética.	Unidad.
gdp	PBI (producto Bruto Interno).	Numérica.	Valor absoluto.
gdp_per_cap	PBI per cápita.	Numérica.	Valor absoluto unitario.
Region	Región.	Alfabética.	Unidad.
Gdp_indicator.	Indicador de PBI.	Numérica.	Unidad.
illiteracy	Analfabetismo.	Numérica.	Valor absoluto.
pop	Población.	Numérica.	Valor absoluto.
poverty	Índice de pobreza.	Numérica.	Valor absoluto.
Deficient_infra	Déficit de infraestructura.	Numérica.	Valor absoluto.
school_dropout	Deserción escolar.	Numérica.	Valor absoluto.

no_healthcare	Falta de acceso a la salud.	Numérica.	Valor absoluto.
birth_mortal	Mortalidad infantil.	Numérica.	Valor absoluto.
movie_theatres_ per_cap	Películas y teatros per cápita.	Numérica.	Valor absoluto unitario.
doctors_per_cap	Doctores per cápita.	Numérica.	Valor absoluto unitario.

Las provincias describen la división geográfica del país formada por 22 elementos. Las regiones por su parte, agrupan un conjunto de provincias en: norte, centro y sur. La población muestra la cantidad de habitantes por provincias. El PBI o Producto Bruto Interno mide la cantidad de bienes y servicios producidos por una economía en un periodo determinado. Al dividir el PBI por la población se obtiene el Producto Bruto Interno per cápita denotado como PBI p/c. El indicador de PBI indica si el PBI p/c de cada provincia es superior o no la media. El analfabetismo expresa la cantidad de personas que no están escolarizadas bajo el sistema educativo formal. El índice de pobreza registra la proporción de personas por provincias que viven en situaciones económicas desfavorables. El déficit de infraestructura muestra la carencia de servicios públicos y de transporte. La deserción escolar indica la tasa de abandono de la escuela obligatoria. La falta de acceso a la salud indica la proporción de la población carente de servicio sanitario. La mortalidad infantil registra la cantidad de infantes que mueren por provincia. Las películas y teatros per cápita, indican el acceso al servicio cultural por habitante en cada provincia. Por último la cantidad de doctores per cápita representa cuanto doctores hay disponible por habitante en cada provincia del país [1] [2] [3].

Capítulo 1

ANALISIS EXPLORATORIO DE DATOS

La primera instancia de un análisis de datos consiste en el análisis exploratorio o preliminar de datos con el fin de evaluar sus características principales y de poder además recabar información básica sobre los datos. Partiendo de la siguiente tabla:

province	gdp	illiteracy	poverty	deficient_infra	school_dropout	no_healthcare	birth_mortal	pop	movie_theatres_per_cap	doctors_per_cap	gdp_per_cap	gdp_indicator	Region
Buenos Aires	292689868	1.383240	8.167798	5.511856	0.7661682	48.7947	4.4	15625084	6.00e-06	0.0048356	18.732051	1	Centro
Catamarca	6150949	2.344140	9.234095	10.464484	0.9519631	45.0456	1.5	367828	5.40e-06	0.0045021	16.722352	0	Norte
Córdoba	69363739	2.714140	5.382380	10.436086	1.0350558	45.7640	4.8	3308876	1.12e-05	0.0101754	20.962931	1	Centro
Corrientes	7968013	5.602420	12.747191	17.438858	3.8642652	62.1103	5.9	992595	4.00e-06	0.0044953	8.027456	0	Norte
Chaco	9832643	7.517580	15.862619	31.479527	2.5774621	65.5104	7.5	1055259	2.80e-06	0.0036048	9.317753	0	Norte
Chubut	17747854	1.548060	8.051752	8.044618	0.5863094	39.5473	3.0	509108	1.57e-05	0.0044981	34.860686	1	Sur
Entre Ríos	20743409	3.185580	7.288751	18.794568	1.8871881	48.6571	3.1	1235994	5.70e-06	0.0046788	16.782775	0	Centro
Formosa	3807057	4.610640	17.035583	28.004984	2.2689741	65.8126	16.2	530162	3.80e-06	0.0034405	7.180932	0	Norte
Jujuy	6484938	2.151390	13.367965	12.483179	0.7212945	54.1615	3.7	673307	3.00e-06	0.0039581	9.631473	0	Norte
La Pampa	6990262	1.539300	3.398774	16.505714	0.2040934	45.4764	7.2	318951	1.88e-05	0.0054146	21.916415	1	Centro
La Rioja	5590516	2.773210	10.875152	7.403254	3.8449494	40.8341	11.4	333642	1.20e-05	0.0050923	16.756031	0	Norte
Mendoza	33431369	2.200200	5.692798	3.839852	1.0637179	50.5843	4.4	1738929	8.10e-06	0.0057202	19.225264	1	Norte
Misiones	9646826	6.863950	13.529788	8.325740	3.1291244	57.8339	8.1	1101593	1.80e-06	0.0028804	8.757160	0	Norte
Neuquén	22564106	1.943750	9.456635	11.267278	1.3935038	48.7431	3.3	551266	9.10e-06	0.0050665	40.931431	1	Sur
Río Negro	10264584	2.031420	8.678391	14.885444	0.4080420	49.9463	0.8	638645	9.40e-06	0.0048979	16.072442	0	Sur
Salta	13438835	3.346090	16.870500	14.182303	1.4820300	60.4230	5.8	1214441	4.10e-06	0.0039911	11.065861	0	Norte
San Juan	8262309	2.963260	9.050784	3.914390	3.2984129	52.9684	4.2	681055	5.90e-06	0.0050436	12.131632	0	Centro
San Luis	11780849	3.433650	6.593771	9.679894	2.0001724	51.6154	3.8	432310	4.60e-06	0.0061021	27.250930	1	Centro
Santa Cruz	11663738	0.791485	8.024762	7.411364	0.2892622	29.2321	3.3	273964	1.10e-05	0.0042706	42.573981	1	Sur
Santa Fe	81588690	1.975940	6.081012	11.869195	2.8721807	41.9660	2.6	3194537	6.60e-06	0.0066717	25.540067	1	Centro
Santiago del Estero	8387859	6.272090	11.759000	20.491433	2.3255981	63.6637	1.7	874006	3.40e-06	0.0028215	9.597027	0	Norte
Tucumán	13856199	3.770370	11.214239	6.466665	0.9772847	48.2242	3.0	1448188	4.80e-06	0.0055007	9.567956	0	Norte

Tabla 1.1. Indicadores socioeconómicos de Argentina.

En la Tabla 1.1. se observa un conjunto de datos formado por 22 provincias y 14 atributos, de los cuales las magnitudes provincias y región son de tipo carácter y las demás variables son de tipo numérico. En primer lugar, con el fin de evaluar que provincias se encuentran por encima del PBI per cápita medio, se crea una variable llamada "*gdp_indicator*", la cual toma un valor igual a uno si el valor del PBI per cápita de la provincia es superior a su media y cero si es inferior. Por otra parte, la variable región permite clasificar las provincias del país en tres áreas geográficas, llamadas: centro, norte y sur. La región centro está formada por las áreas metropolitanas, pampeana y cuyo, mientras que la región sur por la patagónica y la región norte por el noroeste y noreste argentino, como se pueden visualizar en el siguiente mapa:

Figura 1.1. Provincias y regiones de la República Argentina.[1]

La región centro está formada por las regiones: metropolitanas, pampeanas y cuyo, integradas por las provincias de: Buenos Aires, Córdoba, Santa Fe, Entre Ríos, La Pampa, Mendoza, San Luis y San Juan. Por otra parte, la zona norte comprende las regiones noroeste, nordeste formadas por: Catamarca, Salta, Tucumán, Formosa, Misiones, Chaco, Corrientes, Jujuy, La Rioja y Santiago del Estero. Por último, la región sur o patagónica engloba las provincias de: Santa Cruz, Chubut, Neuquén y Rio Negro.

A continuación se muestra el siguiente análisis descriptivo para obtener resultados preliminares sobre la naturaleza de los datos, mediante medidas de centralización, posición y dispersión [4] [5]:

[1] https://www.thinglink.com/scene/509905300401160194

province	gdp	illiteracy	poverty	deficient_infra	school_dropout	no_healthcare	birth_mortal	pop	movie_theatres_per_cap	doctors_per_cap	gdp_per_cap	gdp_indicator
Length:22	Min. : 3807057	Min. :0.7915	Min. : 3.399	Min. : 3.84	Min. :0.2041	Min. :29.23	Min. : 0.800	Min. : 273964	Min. :1.816e-06	Min. :0.002821	Min. : 7.181	Min. :0.0000
Class :character	1st Qu.: 8041587	1st Qu.:1.9898	1st Qu.: 7.473	1st Qu.: 7.57	1st Qu.:0.8126	1st Qu.:45.55	1st Qu.: 3.025	1st Qu.: 514372	1st Qu.:4.052e-06	1st Qu.:0.004061	1st Qu.: 9.606	1st Qu.:0.0000
Mode :character	Median : 10964161	Median :2.7437	Median : 9.142	Median :10.87	Median :1.4378	Median :49.37	Median : 4.000	Median : 777530	Median :5.768e-06	Median :0.004757	Median :16.739	Median :0.0000
NA	Mean : 30557028	Mean :3.2255	Mean : 9.926	Mean :12.68	Mean :1.7249	Mean :50.77	Mean : 4.986	Mean : 1686352	Mean :7.144e-06	Mean :0.004894	Mean :18.346	Mean :0.4091
NA	3rd Qu.: 19994520	3rd Qu.:3.6862	3rd Qu.:12.500	3rd Qu.:16.10	3rd Qu.:2.5145	3rd Qu.:56.92	3rd Qu.: 5.875	3rd Qu.: 1230606	3rd Qu.:9.314e-06	3rd Qu.:0.005334	3rd Qu.:21.678	3rd Qu.:1.0000
NA	Max. :292689868	Max. :7.5176	Max. :17.036	Max. :31.48	Max. :3.8643	Max. :65.81	Max. :16.200	Max. :15625084	Max. :1.881e-05	Max. :0.010175	Max. :42.574	Max. :1.0000

Tabla 1.2. Análisis preliminar de la base de datos.

Las magnitudes con mayor rango de datos son el PBI y la población. El PBI per cápita medio es de 18,346, su valor mínimo de 7,181 y su valor máximo de 42,574. La mediana de la pobreza y del déficit de infraestructura son similares, correspondientes a 9,142 y 10,87 respectivamente. La falta de acceso a la salud presenta un alto grado de dispersión de los datos al igual que el índice de mortalidad infantil y deserción escolar. En los próximos gráficos se pueden analizar cómo inciden ciertos indicadores en cada provincia.

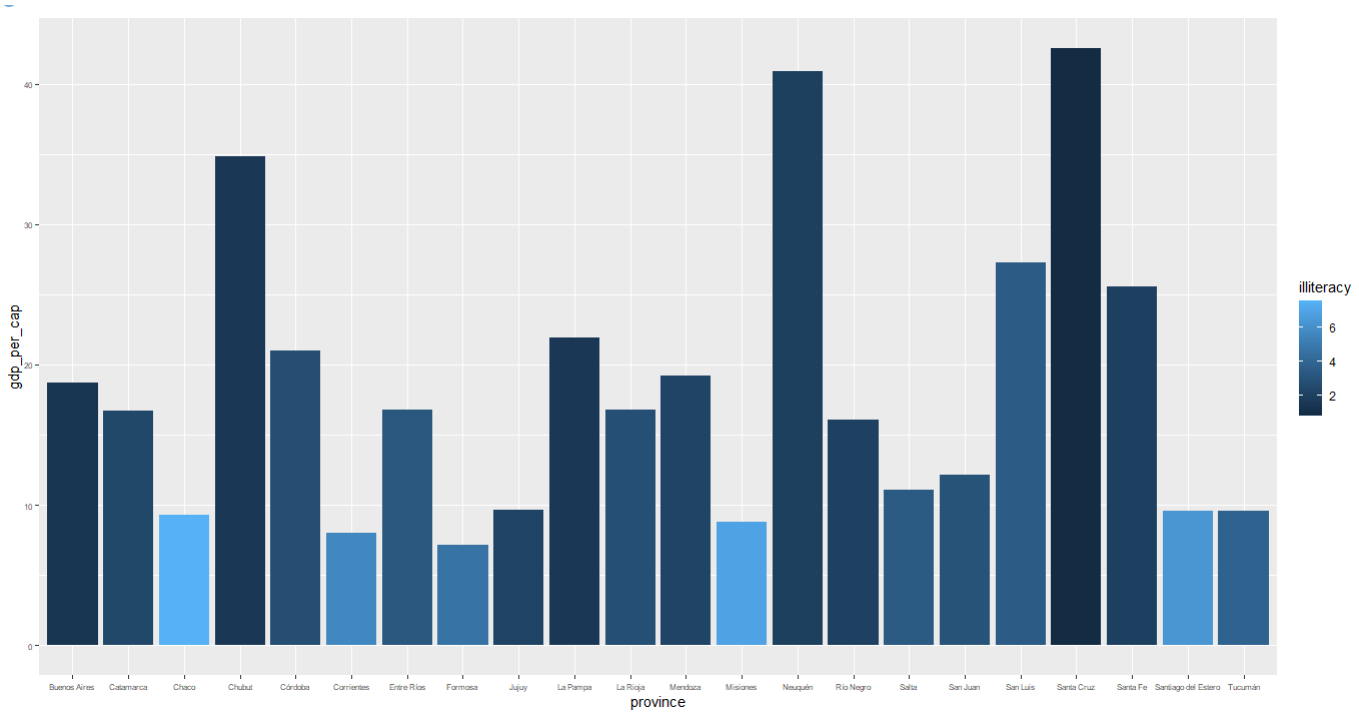


Figura 1.2. Distribución del PBI per cápita y del analfabetismo por provincias.

A partir de la Figura 1.2. se observa que las provincias con mayor PBI per cápita son: Santa Cruz, Neuquén y Chubut, las cuales poseen un bajo índice de analfabetismo, en contraste con Formosa, Corrientes y Misiones, cuyo PBI per cápita es muy bajo y con prevalencia de altos indicadores de analfabetismo.

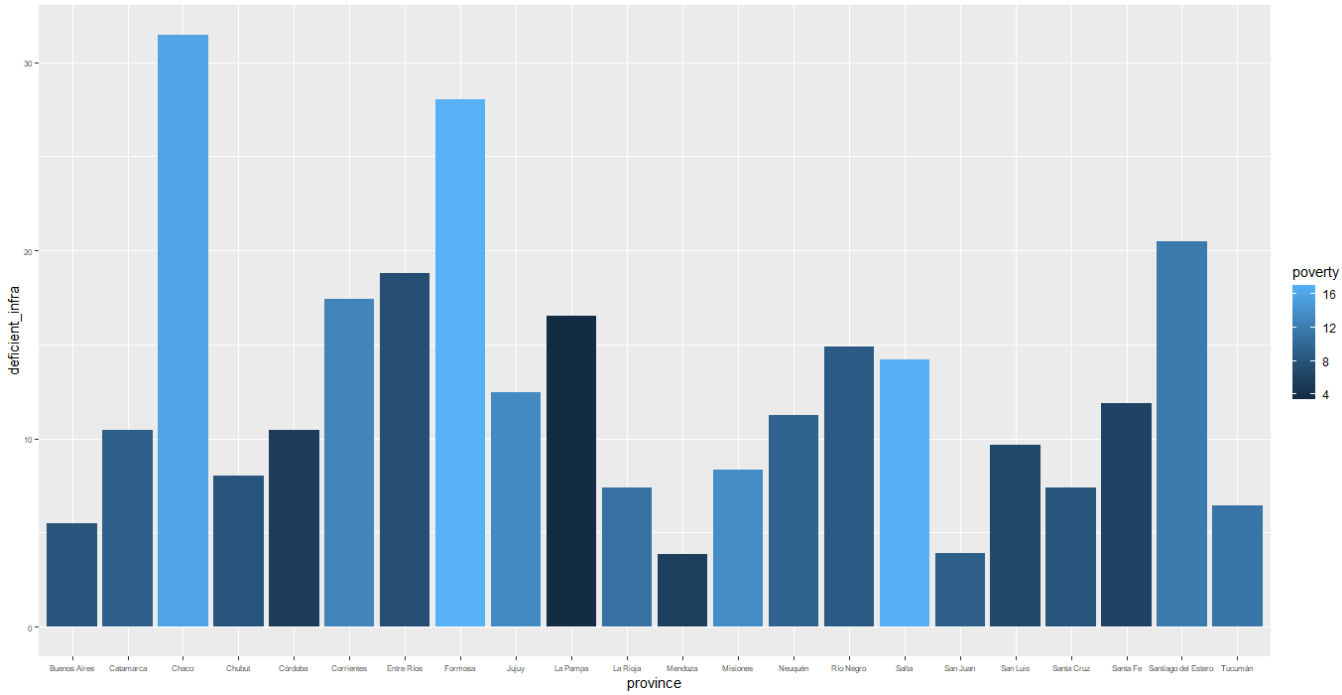


Figura 1.3. Distribución de la pobreza y la falta de infraestructura por provincias.

En la Figura 1.3. se puede apreciar que el índice de pobreza es mayor en Chaco, Formosa y Salta, donde prevalece un gran déficit de infraestructura. Por otra parte, Buenos aires, Mendoza y San Juan representan las provincias con bajo nivel de pobreza y de infraestructura.

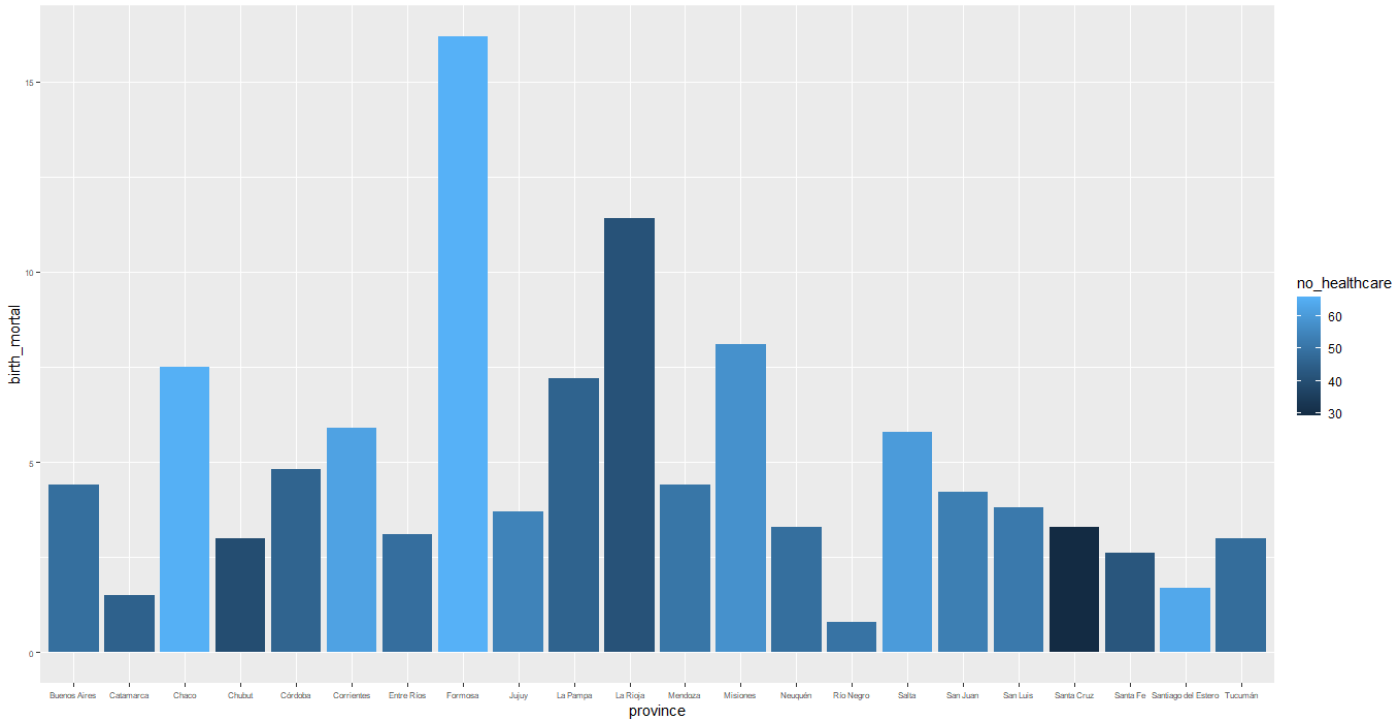


Figura 1.4. Índice de mortalidad infantil y falta de acceso a la salud por provincias.

De acuerdo a la Figura 1.4. las provincias con mayor mortalidad infantil son: Formosa, La Rioja, Misiones y Chaco, donde además prevalece la falta de acceso al sistema de salud. Por otro lado, las provincias con mejores acceso al sistema sanitario y con bajos indicadores de mortalidad infantil son: Santa Cruz, Chubut y Neuquén.

A modo de resumen, se representan: un histograma de frecuencia, un gráfico de dispersión y el coeficiente de correlación [6] [7] de cada par de variables numéricas a través del siguiente gráfico de panel:

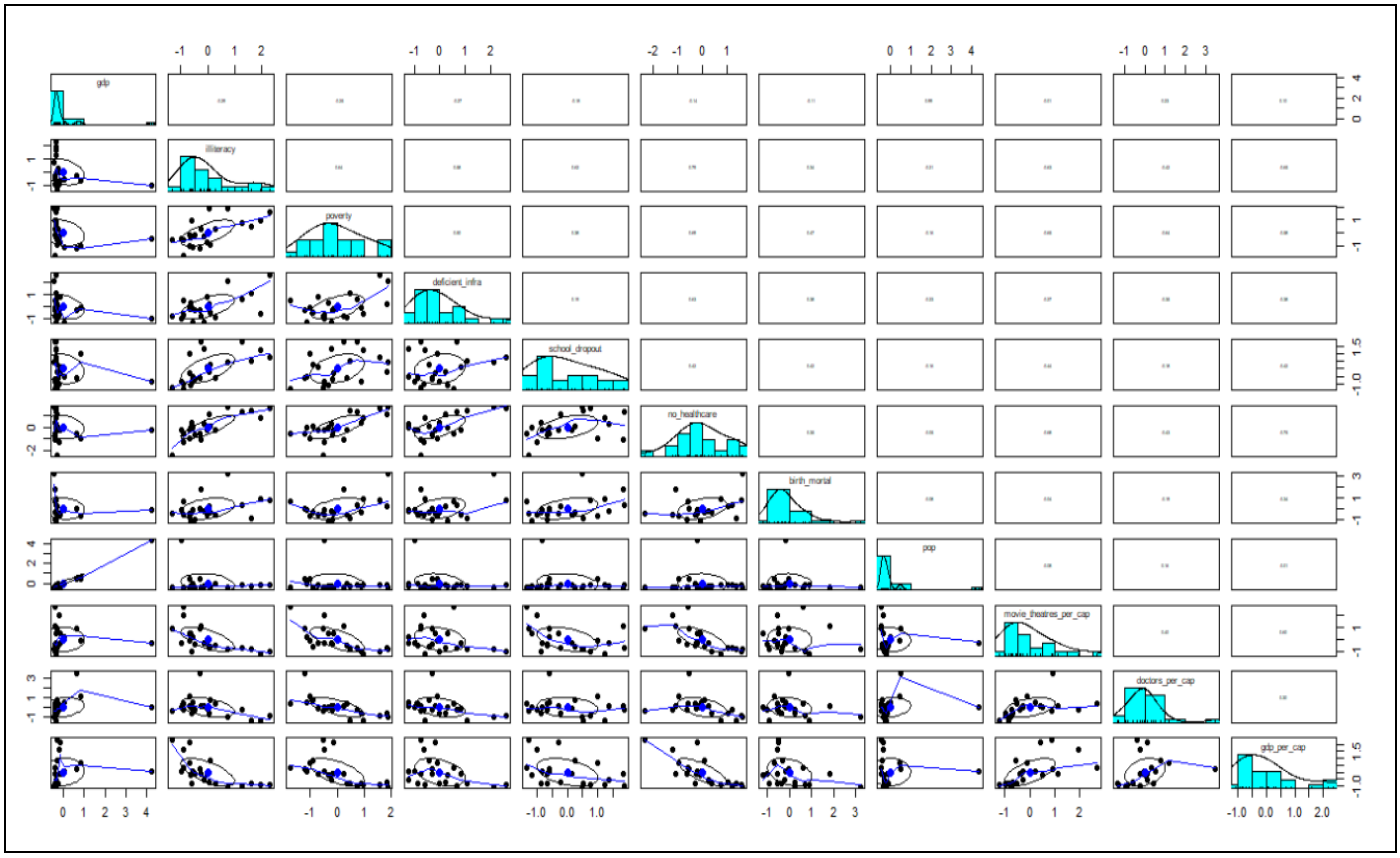


Figura 1.5. Gráfico de panel.

En general se puede notar que en la Figura 1.5. los datos no presentan distribución normal, dado que los valores están sesgados en mayor medida hacia la derecha con excepción de la variable que mide la falta de acceso a la salud, la cual está segada hacia la izquierda.

Capítulo 2

PRUEBAS PARAMÉTRICAS

La prueba paramétrica ANOVA [9] [10] se utiliza para comparar la media de más de dos grupos diferentes, donde el investigador se pregunta si la media es la misma para todos los grupos. En este caso se pretende determinar si la media del PBI per cápita no es la misma para las distintas regiones del país. En primer lugar, se muestra una tabla con valores sobre la media y el desvío standard de cada región:

Region	variable	n	mean	sd
Centro	gdp_per_cap	7	20.474	5.167
Norte	gdp_per_cap	11	11.441	4.106
Sur	gdp_per_cap	4	33.610	12.153

Tabla 2.1. Distribución del PBI per cápita por regiones.

De acuerdo a la Tabla 2.1. se tiene que la región sur posee mayor valor promedio de PBI p/c, seguida de la región centro y norte. Esta relación se puede visualizar a partir del siguiente gráfico:

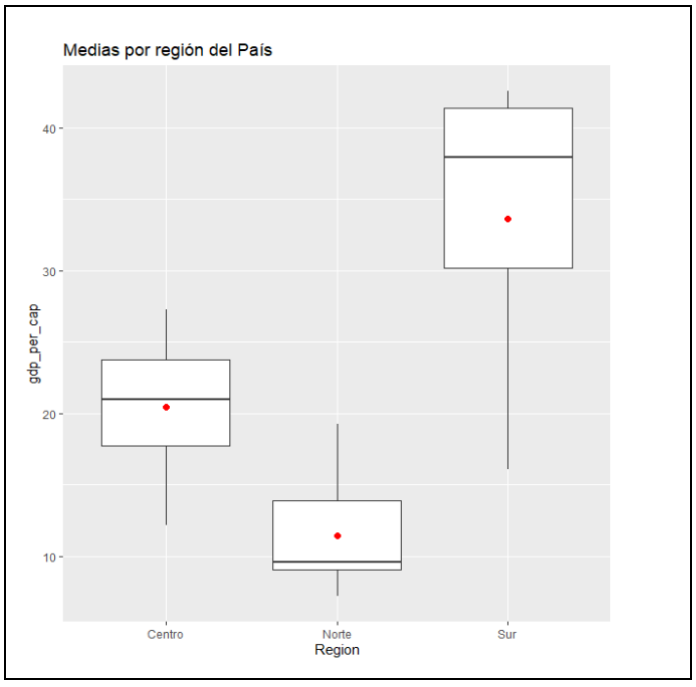


Figura 2.1. Boxplot PBI p/c por regiones.

En la Figura 2.1. se tiene que la mediana del PBI p/c es mucho mayor en la región sur, media en la región centro y baja en la región norte. De acuerdo a la configuración de los datos se puede aplicar un análisis ANOVA para muestras independientes [11] [12], en base al planteo de las siguientes hipótesis:

- Ho*: La media es la misma para todos los grupos.*
- H1*: Al menos uno de los grupos presenta una media diferente.*

Para aplicar la prueba se deben evaluar los supuestos de: normalidad, homocedasticidad y ausencia de outliers. Según el algoritmo de R, los datos presentan ausencia de outliers. Luego, con el fin de evaluar si los datos de distribuyen normalmente a través de cada categoría, se aplica el test de normalidad de Shapiro [13] [14] y se obtienen los siguientes resultados:

Region	variable	statistic	p
Centro	gdp_per_cap	0.9792957	0.9560498
Norte	gdp_per_cap	0.8111927	0.0132067
Sur	gdp_per_cap	0.8340067	0.1784919

Tabla 2.2. Test de normalidad por regiones

De acuerdo a la Tabla 2.2. , el p-valor<5% para la región Norte, por lo tanto la distribución no es normal para este grupo, mientras que el p-valor>5% para las regiones Centro y Sur, donde se observa que los datos presentan distribución normal. Otra forma de aplicar el supuesto de normalidad es a partir de la evaluación de los residuos, mediante el test de normalidad de Shapiro-Wilk, cuyos resultados son:

```
        Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.91441, p-value = 0.0583
```

Tabla 2.3. Test de normalidad de Shapiro-Wilk.

Como e p-valor<5%, se rechaza la Ho, por lo tanto los residuos no se distribuyen normalmente. Por último se aplica el test de Levene para determinar la homogeneidad de la varianza, donde se obtiene:

df1	df2	statistic	p
2	19	1.879599	0.1799567

Tabla 2.4 Test de Levene de homogeneidad de la varianza.

Dado que el p-valor>5% no podemos rechazar la Ho sobre homogeneidad de varianza.

Visto que se comprueba la mayoría de los supuestos, se aplica a continuación la prueba *ANOVA* [15] [16], para determinar si existen diferencias significativas en cada grupo con respecto a la media de la variable PBI per cápita:

Effect	DFn	DFd	F	p	p<.05	ges
Region	2	19	18.316	3.69e-05	*	0.658

Tabla 2.5. Prueba ANOVA.

En la Tabla 2.5. se puede notar que el p-valor<5% , entonces rechazamos la Ho de igualdad entre la media de los grupos, por lo tanto, podemos asegurar que existen diferencias significativas y al menos uno de estos grupos se diferencia del resto con respecto al valor medio de PBI per cápita. Estos resultados se pueden visualizar en el siguiente gráfico:

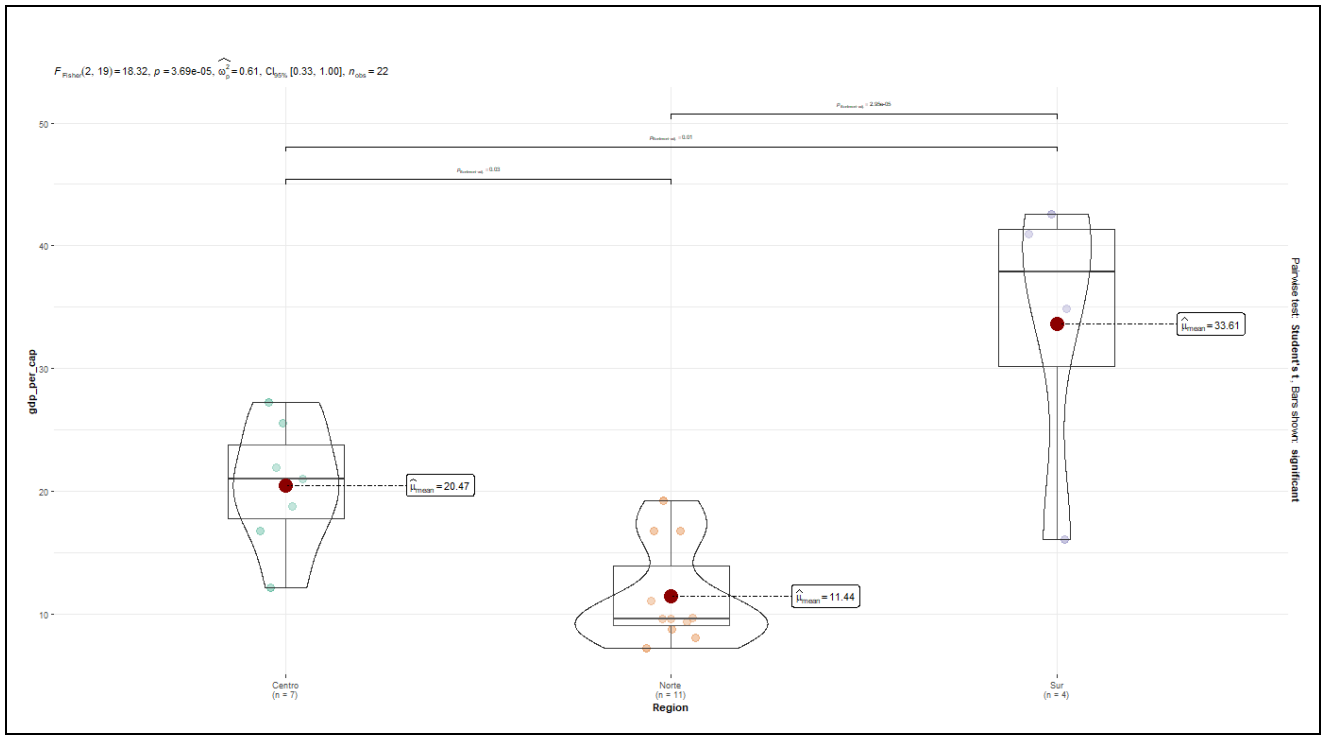


Figura 2.2. Prueba ANOVA.

En relación a la Figura 2.2. se concluye que las medias del PBI per cápita es mayor para la región sur, media para la región centro y menor para la región norte. Como el p-valor<5% existen diferencias significativas entre todos los grupos. La menor diferencia se observa entre las regiones centro y norte, mientras que la mayor variación se da entre los grupos norte y sur.

Capítulo 3

MODELOS DE REGRESIÓN AVANZADOS.

3.1. Modelo de Regresión Lineal Múltiple.

El modelo permite aproximar linealmente más de un predictor con respecto a una variable respuesta, cuya ecuación de aproximación se ve reflejada en un hiperplano [17] [18] [19]. Es decir,

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon\,, \qquad (1)$$

Siendo $\boldsymbol{\beta_o, \beta_1, \beta_2, \ldots, \beta_n}$ los coeficientes de la ecuación del hiperplano, $\boldsymbol{Y}$ la variable respuesta y $\boldsymbol{\varepsilon}$ indica el error o residuo cometido en la aproximación. Al aplicar en modelo de regresión lineal a todos los predictores, considerando al PBI p/c como la variable respuesta se obtienen los siguientes resultados:

```
Call:
lm(formula = gdp_per_cap ~ ., data = Arg_1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6582 -2.2436 -0.8989  1.7899  9.0327

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.896e+01  1.800e+01   2.164  0.05334 .
gdp                      9.987e-07  2.707e-07   3.689  0.00357 **
illiteracy               1.851e+00  1.634e+00   1.133  0.28144
poverty                  3.505e-01  7.256e-01   0.483  0.63859
deficient_infra         -2.790e-01  2.995e-01  -0.932  0.37152
school_dropout          -2.668e+00  1.851e+00  -1.441  0.17738
no_healthcare           -3.095e-01  3.067e-01  -1.009  0.33465
birth_mortal            -7.246e-02  5.554e-01  -0.130  0.89854
pop                     -1.893e-05  5.124e-06  -3.695  0.00353 **
movie_theatres_per_cap   4.259e+05  5.634e+05   0.756  0.46557
doctors_per_cap         -1.529e+03  1.184e+03  -1.292  0.22289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.961 on 11 degrees of freedom
Multiple R-squared:  0.827,     Adjusted R-squared:  0.6698
F-statistic:  5.26 on 10 and 11 DF,  p-value: 0.005631
```

Tabla 3.1. Selección de predictores mediante el método Akaike.

En base a la Tabla 3.1. en la segunda columna se muestran los *coeficientes de cada predictor*, seguidos del *error standard* y del *p-valor* en la última columna, la cual representa el nivel de significación de cada variable observable. En particular se tiene que las variables *PBI* y *población* son estadísticamente significativas, mientras que los demás predictores son poco representativos para el modelo. Además se utilizan las siguientes métricas para medir la eficiencia del modelo: el *error standard residual*, el *coeficiente de determinación ajustado* $\boldsymbol{R^2}$ y el *coeficiente de determinación ajustado* $\boldsymbol{R^2}$ *múltiple*.

El error standard residual representa la diferencia entre los valores observados y predichos por el modelo, mientras que el coeficiente de determinación ajustado $\boldsymbol{R^2}$ mide el porcentaje de la variable respuesta que ha sido explicada por el modelo. Los valores próximos a uno indican mejor desempeño, mientras que los cercanos a cero muestran menor performance del modelo. Por último, el coeficiente de determinación ajustado $\boldsymbol{R^2}$ múltiple es otra forma de evaluar la eficiencia del modelo y su valor varía entre 0 y 1. Los valores cercanos a 1 reflejan un mayor performance y aquellos tendientes a cero menor desempeño del modelo [20] [21].

Para poder determinar qué conjunto de predictores representan la menor perdida de información, se utiliza a continuación el *método de selección automática stepwise* [22].

```
Call:
lm(formula = gdp_per_cap ~ gdp + school_dropout + no_healthcare +
    pop + doctors_per_cap, data = Arg_1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7105 -2.1624 -0.0383  1.8830 10.2896

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.814e+01  1.028e+01   4.684 0.000249 ***
gdp              8.232e-07  1.981e-07   4.155 0.000745 ***
school_dropout  -1.659e+00  1.135e+00  -1.462 0.163202
no_healthcare   -3.600e-01  1.850e-01  -1.947 0.069372 .
pop             -1.573e-05  3.751e-06  -4.194 0.000688 ***
doctors_per_cap -1.489e+03  9.240e+02  -1.611 0.126706
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.34 on 16 degrees of freedom
Multiple R-squared:  0.7981,    Adjusted R-squared:  0.735
F-statistic: 12.65 on 5 and 16 DF,  p-value: 4.336e-05
```

Tabla 3.2. Método de selección automática stepwise.

El método stepwise muestra la combinación de variables que minimizan la perdida de información. Para la variable respuesta PBI p/c, los predictores óptimos son: el PBI, deserción escolar, población, falta de acceso a la salud, médicos p/c. Aplicamos el modelo de regresión lineal múltiple, utilizando los factores seleccionados se obtiene:

A partir de este modelo, se obtiene la siguiente ecuación:

$$Y = \beta_0 + \beta_1 . Gdp + \beta_2 . School_{\text{dropout}} + \beta_3 . \text{No}_{\text{healthcare}} + \beta_4 . \text{Pop} + \beta_5 . Doctors_{\text{per\_cap}} \quad (2)$$

Al reemplazar los coeficientes obtenidos en la Tabla.3.2. se obtiene:

$$\boldsymbol{PBI\ p/c} = 4.814e{+}01 + 8.232e{-}07 . Gdp - 1.659e{+}00 . School_{\text{dropout}} - 3.600e{-}01 . \text{No}_{\text{healthcare}} - 1.573e{-}05 . \text{Pop} - 1.489e{+}03 . Doctors_{\text{per\_cap}} \quad (2b)$$

Ejemplo 1.

Considerando los sig. valores de los predictores:

- Gdp=292689868
- School_dropout=0.7661682
- no_healthcare=48.7947
- Pop=15625084
- doctors_per_cap=0.004835622

Al reemplazar estos resultados en la Fórmula (2b) El valor del PBI p/c es de: **17.96933**, cuyo resultado se aproxima al PBI p/c de la provincia de Buenos Aires.

Observaciones del modelo de regresión (método Stepwise)

Los valores de los coeficientes $\boldsymbol{\beta_2}$, $\boldsymbol{\beta_3}$ y $\boldsymbol{\beta_5}$ presentes en la Tabla 3.2. no son estadísticamente significativos, mientras que $\boldsymbol{\beta_0}, \boldsymbol{\beta_1}$ y $\boldsymbol{\beta_4}$ correspondientes al intercepto y a los predictores *PBI* y *población*, los cuales tienen significación estadística, dado que sus p-valores son inferiores a 0.005. El coeficiente de determinación ajustado $\boldsymbol{R^2}$ es de 0.735. Visto que su valor es próximo a 1, el porcentaje de la variable respuesta que ha sido explicado por el modelo es alta. Entonces, para un nivel de confianza del 95%, dado que el p-valor<5%, se observa que el modelo de regresión lineal múltiple es estadísticamente significativo.

Al aplicar la prueba F, se evalúa si al menos uno de los predictores está significativamente relacionado linealmente con la variable respuesta, es decir, su coeficiente es distinto de cero. Dado que su valor es diferente de cero se comprueba que los predictores no son colineales.

Luego, al evaluar el R^2parcial, se puede determinar el porcentaje de la variable respuesta que será explicada al agregar un predictor en particular. Se calcula elevando al cuadrado el coeficiente de correlación parcial entre el predictor y la variable independiente del modelo, cuyos resultados son:

```
$adjustment
[1] FALSE

$variable
[1] "gdp"             "school_dropout"  "no_healthcare"   "pop"             "doctors_per_cap"

$partial.rsq
[1] 0.5190228 0.1177959 0.1914652 0.5236170 0.1395837
```

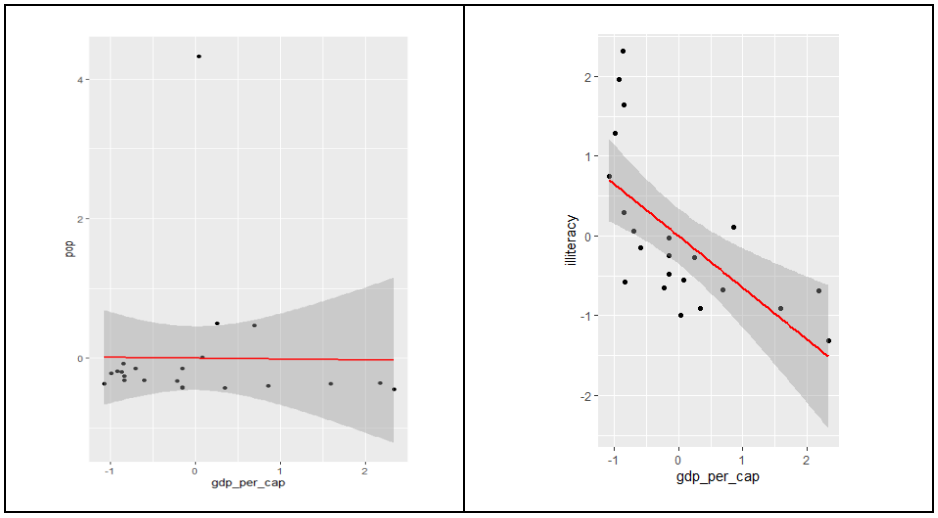
Tabla 3.3. Valores del R^2parcial

En base a la Tabla 3.3. se puede notar que el R^2parcial de la variable *deserción escolar* es muy bajo, por lo tanto el porcentaje de la variable respuesta explicada por esta magnitud es poco significativo. Por otra parte, el R^2 del predictor *población* es muy alto y cercano a 1, lo cual comprueba que su comportamiento es muy influyente para el modelo.

3.1.1. Supuestos del modelo de regresión lineal múltiple

Para poder implementar un modelo de regresión lineal múltiple se deben cumplir los supuestos de: linealidad, homocedasticidad y normalidad.

Supuesto de linealidad



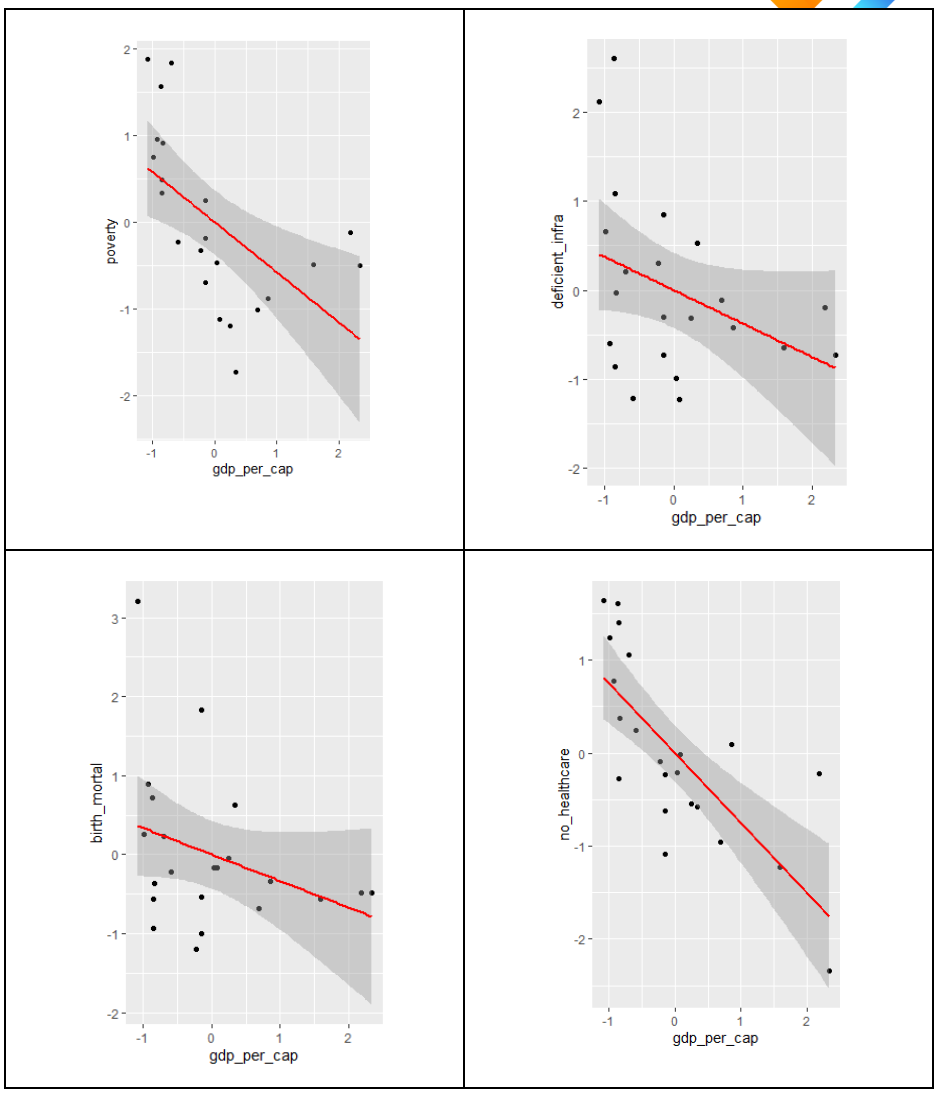


Figura 3.1. Linealidad de los predictores

En todos los gráficos de dispersión se puede observar la relación lineal entre cada predictor y la variable respuesta.

Análisis de los residuos

Los siguientes gráficos de residuos permiten evaluar los supuestos del modelo de regresión lineal múltiple [23] [24].

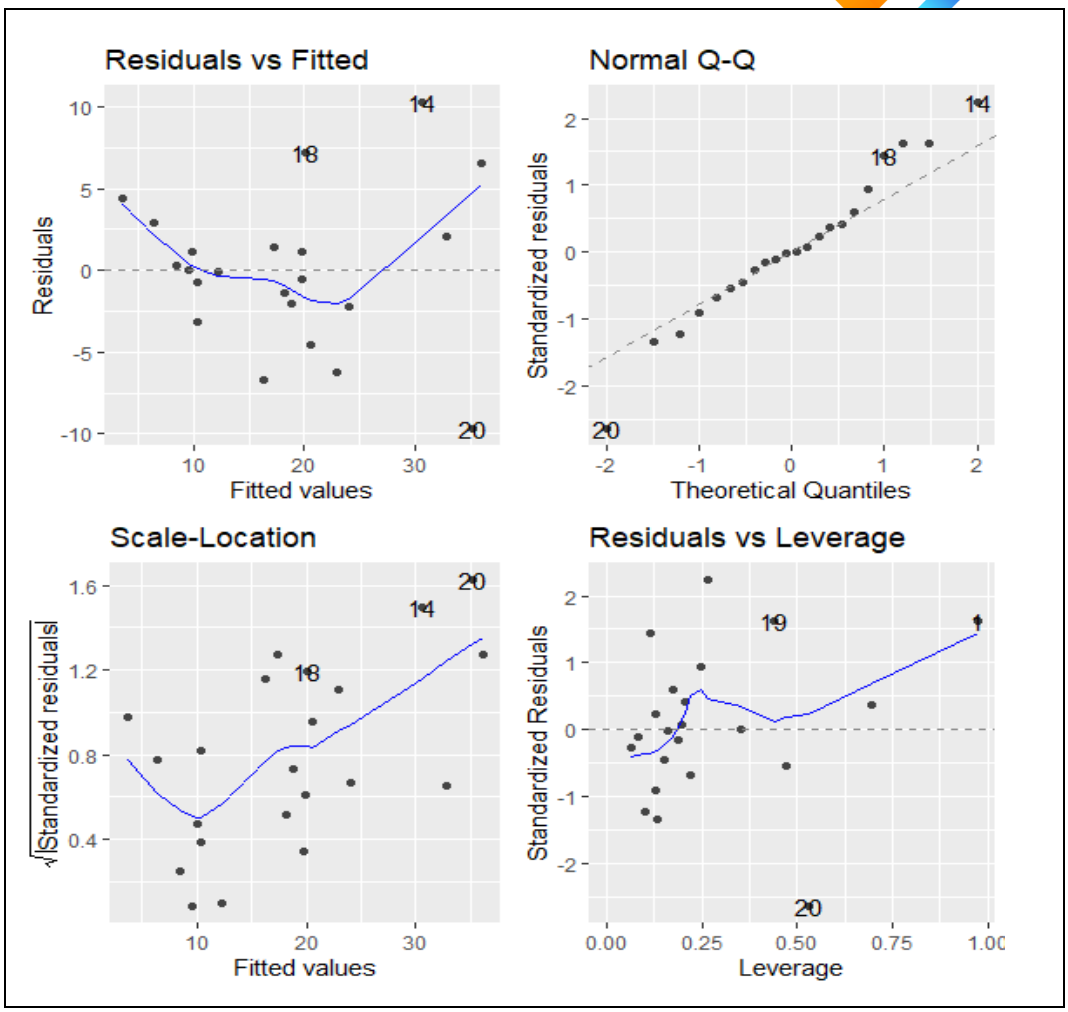


Figura 3.2. Gráfico de residuos.

La figura "Residuals vs Fitted", muestra el gráfico de los residuos en función de los valores ajustados por el modelo y permite evaluar los supuestos de linealidad y homocedasticidad. En este caso es casi lineal, pero es heterocedástico. El gráfico "Scale-Location", muestra la distribución de los datos. Utiliza en valor absoluto la raíz cuadrada de los residuos estandarizados y los representa en función de los valores ajustados por el modelo. Permite evaluar el supuesto de homocedasticidad. En esta representación no se observa una distribución uniforme de los residuos. Por otra parte, La imagen "Normal Q-Q", también llamado gráfico cuantil-cuantil normal, compara los cuantiles de los datos con los cuantiles teóricos de la distribución normal estándar. Permite evaluar el supuesto de normalidad, el cual se verifica dado que los puntos se alinean a la diagonal.

Por último, el gráfico "Residuals vs Leverage", muestra los residuos estandarizados en función de los valores de apalancamiento. Indica también contornos de la misma distancia de Cook (por defecto, 0.5 y 1). Permite evaluar si existen valores atípicos, de apalancamiento o influyentes,

que pueden perturbar los resultados del modelo. En esta instancia se observa la presencia de tres valores influyentes: 14,16 y 20 que se alejan de la curva.

3.2. Modelo de Regresión Logística.

El modelo de regresión logística evalúa la probabilidad de ocurrencia de un evento donde la variable respuesta es categórica dicotómica, es decir , admite dos posibles respuestas, como por ejemplo: si o no, verdadero o falso, aumenta o disminuye, etc. Gráficamente se representa a través de la siguiente función, llamada sigmoide [25] [26].

$$Y = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}, \quad (3)$$

Continuando con el análisis de la influencia del PBI p/c sobre las demás variables y regiones, supongamos que se quiere determinar cuál es la probabilidad que una provincia presente un valor de PBI p/c superior a su media. A partir de los resultados presentes en la magnitud "gdp_indicator", creada al comienzo del informe en el Capítulo.1, página 7, se evalúa cuál es la probabilidad de que las provincias tengan un valor de PBI p/c alto.
Entonces, se genera un modelo de regresión logística, considerando las variables mencionadas y se obtiene:

```
Call:  glm(formula = gdp_indicator ~ gdp_per_cap, family = "binomial",
    data = Arg)

Coefficients:
(Intercept)  gdp_per_cap
    -383.61        21.58

Degrees of Freedom: 21 Total (i.e. Null);  20 Residual
Null Deviance:      29.77
Residual Deviance: 3.962e-09    AIC: 4
```

Tabla 3.4. Modelo de regresión logística.

A partir de la Tabla 3.4. los valores del intercepto y del coeficiente son respectivamente: -383.61 y 21.58 . Luego, con el fin de obtener información detallada del modelo, se muestra la siguiente información:

```
Call:
glm(formula = gdp_indicator ~ gdp_per_cap, family = "binomial",
    data = Arg)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.232e-05  -2.100e-08  -2.100e-08   2.100e-08   4.524e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -383.61  246914.35  -0.002    0.999
gdp_per_cap    21.58   13864.84   0.002    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.9767e+01  on 21  degrees of freedom
Residual deviance: 3.9621e-09  on 20  degrees of freedom
AIC: 4

Number of Fisher Scoring iterations: 25
```

Tabla 3.5. Resumen estadístico del modelo de regresión logística.

Los datos se distribuyen normalmente dado que los valores del 1er y 3er cuartil son similares en valor absoluto, al igual que el módulo de los valores máximos y mínimos. Tanto el intercepto como el coeficiente no son estadísticamente significativos. En base a la Ecuación 3 y a los datos de la Tabla 3.5. se construye la función del modelo de regresión logística:

$$Y = \frac{e^{(\beta_0 + \beta_1 x_1)}}{1 + e^{(\beta_0 + \beta_1 x_1)}}. \quad (4)$$

$$Y = \frac{e^{(-383.61 + 21.58 * \text{gdp\_per\_cap})}}{1 + e^{(-383.61 + 21.58 * \text{gdp\_per\_cap})}}. \quad (5)$$

Ejemplo 2.

Al tomar un valor del PBI p/c de=18.732051, se pretende determinar cuál es la probabilidad que una provincia presente un valor que se encuentre por encima de su media.

Entonces, al reemplazar este valor en la Ecuación (5), se obtiene:

$$Y = \frac{e^{(-383.61 + 21.58 * 18.732051)}}{1 + e^{(-383.61 + 21.58 * 18.732051)}}. \quad (6)$$

Y= 1, cuyo resultado representa una muy alta probabilidad que la provincia tenga un PBI p/c superior a la media.

Para evaluar la posibilidad de que el evento ocurra, se calcula el *odd ratio*. Como el *odd ratio*>1 entonces aumenta la posibilidad de que el evento ocurra.

3.2.1. Supuestos del modelo de regresión logística

Con la finalidad de evaluar los supuestos que permiten aplicar un modelo de regresión logística, se muestran a continuación el siguiente gráfico de residuos [27]:

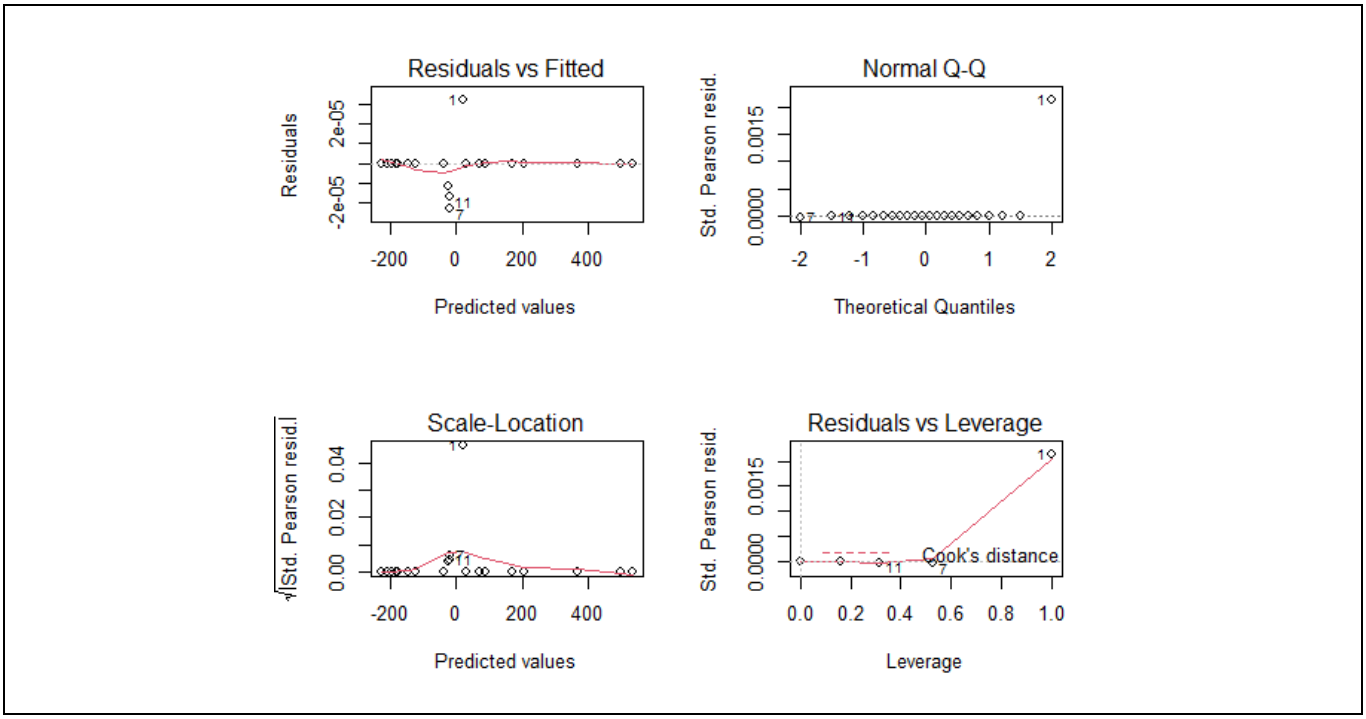


Figura 3.3. Gráfico de residuos modelo de regresión logística.

De acuerdo al grafico Q-Q se observa que la distribución de los residuos es normal, no obstante, en el gráfico de Cook se observa la presencia de valores influentes en los extremos de las curvas.

A continuación se determina la probabilidad de que el evento ocurra de acuerdo al valor del predictor *gdp_per_cap*:

```
 model: gdp_indicator ~ gdp_per_cap

 gdp_per_cap effect
gdp_per_cap
         7.2           16           25           34           43
2.220446e-16 2.220446e-16 1.000000e+00 1.000000e+00 1.000000e+00
```

Tabla 3.6. Probabilidad de ocurrencia del PBI p/c.

A partir de la Tabla 3.6. se observa que cuando el valor del PBI p/c de una provincia aumenta, se incrementa la probabilidad que su valor sea superior a la media. Por ejemplo, para una provincia cuyo PBI p/c es de 7.2 la probabilidad de superar a la media es muy baja, mientras que si su valor es de 43 existe mayor probabilidad que dicho valor supere al PBI p/c promedio.

Ejemplo 3.

Considerando los siguientes valores de PBI p/c= 18.732051, 16.722352,20.962931 ,8.027456, 9.317753, tomados de la Tabla 1.1., al aplicar la función (5) la probabilidad de que el valor supere el PBI p/c medio es:

	PBI_PC	Prob
1	18.732051	1.000000e+00
2	16.722352	1.328708e-10
3	20.962931	1.000000e+00
4	8.027456	4.307377e-92
5	9.317753	5.333017e-80

Tabla 3.7. Probabilidad en función del PBI p/c.

De acuerdo a la Tabla 3.7. se observa que la probabilidad de presentar superávit es alta para el primer registro y baja para las demás observaciones. Con el fin de visualizar estas observacion es, se muestra la siguiente representación de la regresión logística:

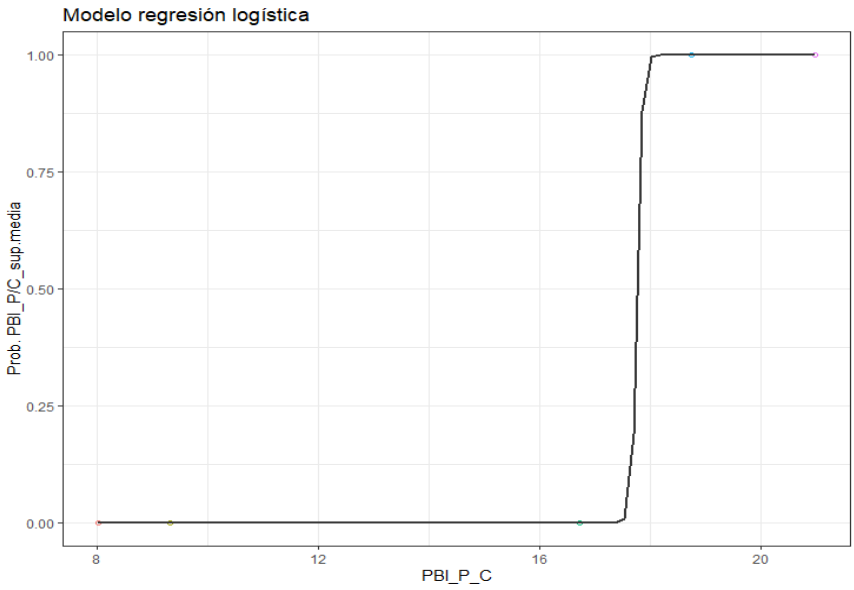


Figura 3.4. Función Sigmoide del modelo de regresión logística.

De acuerdo a la Figura 3.4. se observa que para valores del PBI_P/C inferiores a 18 presentan una baja probabilidad que su valor se encuentre superior a la media.

Observaciones del modelo de regresión logística.

El modelo se utiliza contemplando sus limitaciones de aplicabilidad, dado que se analiza un solo año y no se contempla la variabilidad.

Capítulo 4

MODELO DE CLÚSTER

El modelo de clúster es un algoritmo de aprendizaje no supervisado, donde los datos no están etiquetados y se pueden agrupar observaciones que presentan características comunes, con el fin de identificar subgrupos homogéneos entre si diferenciables del resto [28] [29].

Entre los diferentes métodos de agrupamiento, se encuentra el clúster jerárquico y no jerárquico. El método no jerárquico necesita que el número de agrupamientos sea determinado de antemano. El método más utilizado se llama k-means, en el cual se calcula la distancia entre el centroide de cada grupo y los elementos próximos. Po otro lado, el método jerárquico, establece una jerarquía de agrupamiento sin necesidad de fijar el número de clústers. Entre sus métodos más utilizados, se encuentran: el vecino más lejano, el vecino más cercano, método de Ward o varianza mínima, el método del centroide y de la mediana [30] [31]. Para aplicar el modelo, en primer lugar se seleccionan las magnitudes y se evalúan sus datos:

variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	se	ci
gdp_per_cap	22	7.181	42.574	16.739	9.606	21.678	12.072	10.563	18.346	10.374	2.212	4.599
poverty	22	3.399	17.036	9.142	7.473	12.500	5.027	3.829	9.926	3.780	0.806	1.676

Tabla 4.1. Resumen estadístico del PBI p/c y del índice de pobreza.

Como los datos están desbalanceados se deben estandarizar:

	Arg.gdp_per_caps	Arg.povertys
Buenos Aires	0.0372465	-0.4650914
Catamarca	-0.1564821	-0.1829670
Córdoba	0.2522961	-1.2020658
Corrientes	-0.9946420	0.7465388
Chaco	-0.8702616	1.5708287
Chubut	1.5919951	-0.4957950
Entre Ríos	-0.1506575	-0.6976723
Formosa	-1.0762442	1.8811752
Jujuy	-0.8400200	0.9107853
La Pampa	0.3442089	-1.7268947
La Rioja	-0.1532355	0.2512289
Mendoza	0.0847906	-1.1199344
Misiones	-0.9243010	0.9536008
Neuquén	2.1771954	-0.1240868
Río Negro	-0.2191313	-0.3299970
Salta	-0.7017496	1.8374970
San Juan	-0.5990127	-0.2314681
San Luis	0.8584390	-0.8815523
Santa Cruz	2.3355319	-0.5029362
Santa Fe	0.6935174	-1.0172195
Santiago del Estero	-0.8433405	0.4850803
Tucumán	-0.8461428	0.3409457

Tabla 4.2. Valores estandarizados del PBI p/c y de la pobreza.

Luego, se analiza nuevamente la distribución de los datos estandarizados y se comprueba que los valores son más homogéneos.

Arg.gdp_per_caps		Arg.povertys	
Min.	:-1.0762	Min.	:-1.7269
1st Qu.	:-0.8425	1st Qu.	:-0.6490
Median	:-0.1549	Median	:-0.2072
Mean	: 0.0000	Mean	: 0.0000
3rd Qu.	: 0.3212	3rd Qu.	: 0.6812
Max.	: 2.3355	Max.	: 1.8812

Tabla 4.3. Estadísticos de valores estandarizados.

Con el fin de evaluar si el conjunto de datos presenta una estructura de agrupación adecuada al método de clúster, se aplica la prueba de Hopkins [32], la cual es una prueba que mide la aleatoriedad de los datos, es decir, evalúa la probabilidad que presenta el conjunto de datos de no contener una estructura de agrupación adecuada, comparando la distancia entre un conjunto de datos aleatorios con el conjunto real.

Si el valor del indicador es igual a 0.5 entonces no hay tendencia de agrupación, caso contrario, la asociación es adecuada. Al aplicar la prueba de Hopkins la tendencia de agrupación es de: 0.968383, lo cual asegura una buen nivel de agrupación de los datos analizados.
Continuando con los procedimientos de clusterización, se determinan las *distancias euclideas* [33] entre los datos:

	1	2	3	4	5	6	7	8	9	10	11
1	0.00000000	0.34223514	0.76770941	1.59149037	2.22902251	1.55505175	0.29900127	2.59708072	1.63175762	1.29860448	0.74121395
2	0.34223514	0.00000000	1.09802642	1.25159628	1.89348373	1.77624145	0.51473818	2.25978882	1.28977446	1.62308475	0.43420807
3	0.76770941	1.09802642	0.00000000	2.31342062	2.99150128	1.51446746	0.64558849	3.35728979	2.37850674	0.53281637	1.50881461
4	1.59149037	1.25159628	2.31342062	0.00000000	0.83362119	2.86950953	1.67273895	1.13756697	0.22557679	2.81254249	0.97636922
5	2.22902251	1.89348373	2.99150128	0.83362119	0.00000000	3.21459818	2.37990055	0.37248328	0.66073585	3.51424503	1.50182222
6	1.55505175	1.77624145	1.51446746	2.86950953	3.21459818	0.00000000	1.75430679	3.57344210	2.80947777	1.75287668	1.89838736
7	0.29900127	0.51473818	0.64558849	1.67273895	2.37990055	1.75430679	0.00000000	2.73992062	1.74995895	1.14201207	0.94890467
8	2.59708072	2.25978882	3.35728979	1.13756697	0.37248328	3.57344210	2.73992062	0.00000000	0.99872843	3.87760950	1.87314441
9	1.63175762	1.28977446	2.37850674	0.22557679	0.66073585	2.80947777	1.74995895	0.99872843	0.00000000	2.89132387	0.95220139
10	1.29860448	1.62308475	0.53281637	2.81254249	3.51424503	1.75287668	1.14201207	3.87760950	2.89132387	0.00000000	2.03971173
11	0.74121395	0.43420807	1.50881461	0.97636922	1.50182222	1.89838736	0.94890467	1.87314441	0.95220139	2.03971173	0.00000000
12	0.65656674	0.96753316	0.18655740	2.15613016	2.85522869	1.63132315	0.48346783	3.21786589	2.23138907	0.66007471	1.39166998
13	1.71384389	1.37161665	2.45586631	0.21868360	0.61958902	2.90387569	1.82352047	0.93993672	0.09453283	2.96549715	1.04300920
14	2.16694839	2.33442010	2.20619032	3.28915513	3.48708079	0.69327221	2.39747763	3.82177248	3.18975682	2.43491936	2.36045968
15	0.28979311	0.15982100	0.99133639	1.32678044	2.00925585	1.81869948	0.37399697	2.37148164	1.38745936	1.50621224	0.58494945
16	2.41826967	2.09274733	3.18577232	1.12959091	0.31544928	3.27192852	2.59437574	0.37703316	0.93697023	3.71468937	1.67842606
17	0.67779464	0.44518054	1.29104095	1.05499767	1.82259421	2.20689461	0.64681431	2.16587431	1.16740193	1.76804066	0.65704923
18	0.92075885	1.23210638	0.68566618	2.46669616	3.00042966	0.82880225	1.02571318	3.37278256	2.46925834	0.98946267	1.51877533
19	2.29859698	2.51247166	2.19741970	3.55685904	3.81806400	0.74357109	2.49380424	4.16223539	3.47602330	2.33740056	2.60052460
20	0.85763457	1.19099805	0.47837677	2.44145977	3.02380528	1.03881938	0.90263041	3.39599002	2.46352175	0.79098369	1.52510720
21	1.29547658	0.95815537	2.01168628	0.30208066	1.08608212	2.62544774	1.37066165	1.41538870	0.42571796	2.51059892	0.72865038
22	1.19585635	0.86609267	1.89405724	0.43192336	1.23011945	2.57772216	1.24997087	1.55732254	0.56987246	2.38598015	0.69869139

	12	13	14	15	16	17	18	19	20	21	22
1	0.65656674	1.71384389	2.16694839	0.28979311	2.41826967	0.67779464	0.92075885	2.29859698	0.85763457	1.29547658	1.19585635
2	0.96753316	1.37161665	2.33442010	0.15982100	2.09274733	0.44518054	1.23210638	2.51247166	1.19099805	0.95815537	0.86609267
3	0.18655740	2.45586631	2.20619032	0.99133639	3.18577232	1.29104095	0.68566618	2.19741970	0.47837677	2.01168628	1.89405724
4	2.15613016	0.21868360	3.28915513	1.32678044	1.12959091	1.05499767	2.46669616	3.55685904	2.44145977	0.30208066	0.43192336
5	2.85522869	0.61958902	3.48708079	2.00925585	0.31544928	1.82259421	3.00042966	3.81806400	3.02380528	1.08608212	1.23011945
6	1.63132315	2.90387569	0.69327221	1.81869948	3.27192852	2.20689461	0.82880225	0.74357109	1.03881938	2.62544774	2.57772216
7	0.48346783	1.82352047	2.39747763	0.37399697	2.59437574	0.64681431	1.02571318	2.49380424	0.90263041	1.37066165	1.24997087
8	3.21786589	0.93993672	3.82177248	2.37148164	0.37703316	2.16587431	3.37278256	4.16223539	3.39599002	1.41538870	1.55732254
9	2.23138907	0.09453283	3.18975682	1.38745936	0.93697023	1.16740193	2.46925834	3.47602330	2.46352175	0.42571796	0.56987246
10	0.66007471	2.96549715	2.43491936	1.50621224	3.71468937	1.76804066	0.98946267	2.33740056	0.79098369	2.51059892	2.38598015
11	1.39166998	1.04300920	2.36045968	0.58494945	1.67842606	0.65704923	1.51877533	2.60052460	1.52510720	0.72865038	0.69869139
12	0.00000000	2.30603862	2.31729799	0.84638622	3.06023626	1.12114202	0.80954175	2.33377868	0.61733184	1.85404952	1.73228401
13	2.30603862	0.00000000	3.28339611	1.46454347	0.91148317	1.22890222	2.55850513	3.57043559	2.54979783	0.47546404	0.61762037
14	2.31729799	3.28339611	0.00000000	2.40515707	3.48369561	2.77828399	1.52081302	0.41060605	1.73175818	3.08135060	3.05889343
15	0.84638622	1.46454347	2.40515707	0.00000000	2.22057439	0.39245104	1.21052508	2.56051011	1.14245457	1.02663924	0.91831781
16	3.06023626	0.91148317	3.48369561	2.22057439	0.00000000	2.07151424	3.13487117	3.83441083	3.17744809	1.35980840	1.50350092
17	1.12114202	1.22890222	2.77828399	0.39245104	2.07151424	0.00000000	1.59586181	2.94707429	1.51262665	0.75705854	0.62348285
18	0.80954175	2.55850513	1.52081302	1.21052508	3.13487117	1.59586181	0.00000000	1.52484540	0.21355269	2.18259892	2.09764166
19	2.33377868	3.57043559	0.41060605	2.56051011	3.83441083	2.94707429	1.52484540	0.00000000	1.72066817	3.32887463	3.29168508
20	0.61733184	2.54979783	1.73175818	1.14245457	3.17744809	1.51262665	0.21355269	1.72066817	0.00000000	2.14914796	2.05308706
21	1.85404952	0.47546404	3.08135060	1.02663924	1.35980840	0.75705854	2.18259892	3.32887463	2.14914796	0.00000000	0.14416180
22	1.73228401	0.61762037	3.05889343	0.91831781	1.50350092	0.62348285	2.09764166	3.29168508	2.05308706	0.14416180	0.00000000

Tabla 4.4. Matriz de distancia euclidea.

A partir de los datos obtenido en la Tabla 4.4. se representan en un mapa de calor las distancias entre los puntos para determinar la cantidad de clúster a utilizar:

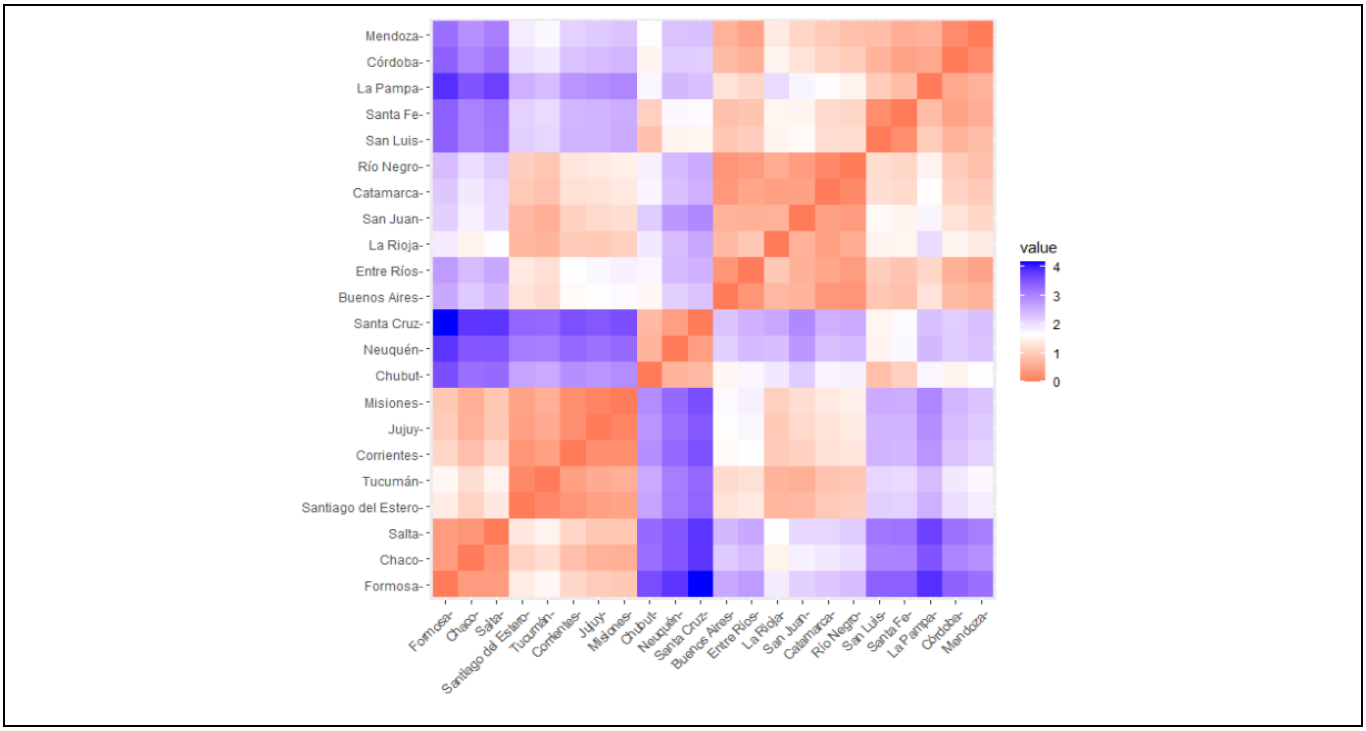


Figura 4.1. Distancias euclideas en un mapa de calor.

En la Figura 4.1., los rectángulos rojos muestran poca diferencia entre los puntos, mientras que los de color azul indican mucha diferencia de distancia entre las observaciones. En el mapa de calor, se visualizan entre 4 y 5 rectángulos rojos, correspondientes a 4 o 5 clústers. Para determinar el número de clúster también se pueden aplicar el método de validación interna, a través de los siguientes gráficos:

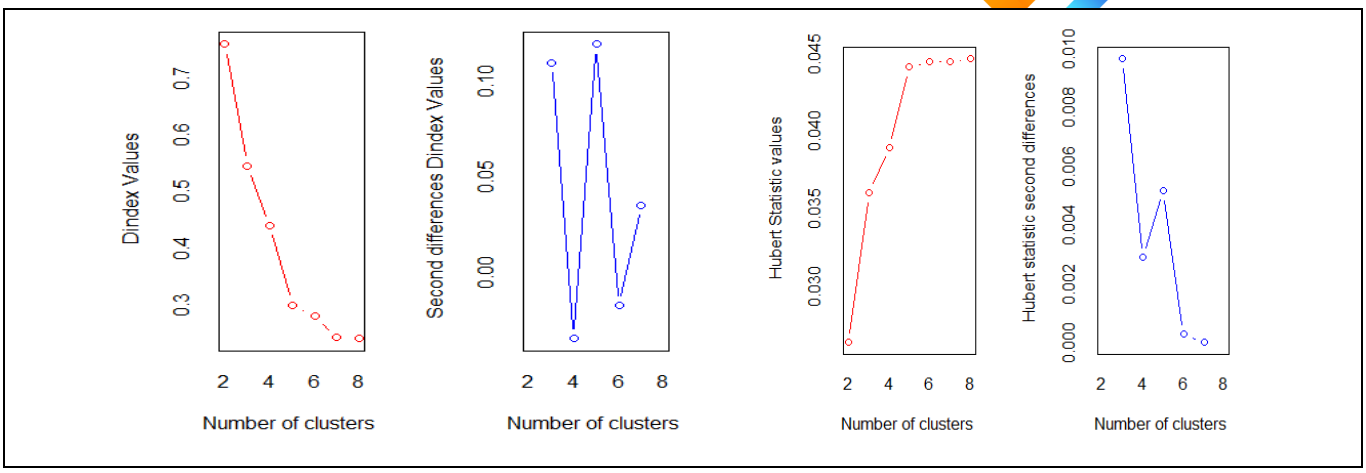


Figura 4.2. Determinación del número de clúster.

En la Figura 4.2. se observa un punto de quiebre de la curva entre 4 y 6 clústers. Además, de acuerdo al *criterio de calidad interna* [34], cuyo método compara treinta estadísticos y determina la cantidad óptima de grupos, en base a la mayoría de las reglas aplicadas, la cantidad óptima de clústers es 5, cuyas características se presentan a continuación:

```
K-means clustering with 5 clusters of sizes 3, 5, 6, 3, 5

Cluster means:
  Arg.gdp_per_caps Arg.povertys
1       -0.8827518    1.7631669
2        0.4466504   -1.1895334
3       -0.2068788   -0.2759945
4        2.0349074   -0.3742727
5       -0.8896893    0.6873902

Clustering vector:
 [1] 3 3 2 5 1 4 3 1 5 2 3 2 5 4 3 1 3 2 4 2 5 5

Within cluster sum of squares by cluster:
[1] 0.1268020 0.8280262 0.7271818 0.4007072 0.3039497
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Tabla 4.5. Características del clúster.

En la Tabla 4.5. se muestran los criterios de: cardinalidad, calidad, magnitud y centroide de cada clúster. La cardinalidad indica la cantidad de elementos en cada clúster cuyos resultados son: 3, 5, 6, 3, 5 para el clúster: 1, 2, 3, 4 y 5 respectivamente. Por otra parte, la calidad de agrupación es muy buena correspondiente a 94.3%. La magnitud indica la variabilidad entre los grupos la cual es muy grande, cuyo valor es: 0.1268020 0.8280262 0.7271818 0.4007072 0.3039497 para cada clúster. Por último, la distancia de cada variable al centroide de cada grupo es:

	Arg.gdp_per_caps	Arg.povertys
1	-0.8827518	1.7631669
2	0.4466504	-1.1895334
3	-0.2068788	-0.2759945
4	2.0349074	-0.3742727
5	-0.8896893	0.6873902

Siendo mayor su valor para la variable PBI p/c para el clúster 4 y menor para el clúster 3, mientras que para el índice de pobreza el valor de la distancia del centroide es mayor para el clúster 1 y menor para el clúster 3. Para visualizar las diferentes agrupaciones de los datos, se muestra el siguiente gráfico:

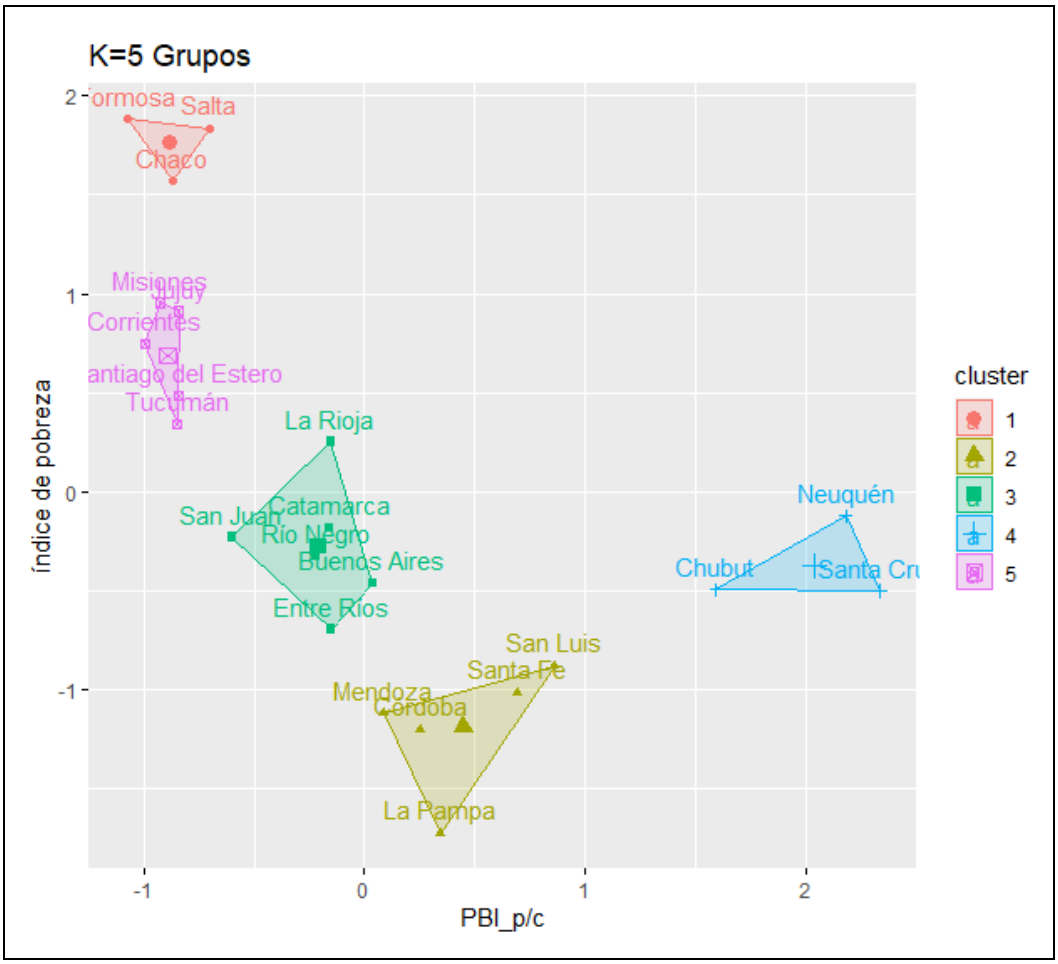


Figura 4.3. Visualización de cada clúster.

El clúster 1, está formado por las provincias de: Chaco, Formosa y Salta con el menor valor de PBI p/c y con el mayor índice de pobreza. Seguido del clúster 2 integrado por: Córdoba, La Pampa, Mendoza, San Luis y Santa Fe cuyos índices de pobreza son los más bajos del país y presentan in valor de PBI p/c medio. Luego el clúster 3 correspondiente a: Buenos Aires, Catamarca, Entre Ríos, La Rioja, Rio Negro y San Juan tienen valores medios de PBI p/c y de pobreza. Por su parte el clúster 4 referente a: Chubut, Neuquén y Santa Cruz presentan PBI p/c muy alto pero con valores de pobreza medios. Por último el clúster 5, correspondientes a:

Corrientes, Jujuy, Misiones, Santiago del Estero y Tucumán presentan altos niveles de pobreza y bajos niveles de PBI p/c. Tanto el clúster 1 representa el mayor índice de pobreza mientras que el clúster 2 el menor. Por otra parte, el grupo 4 presenta mayor PBI p/c mientras que el clúster 1 posee el menor índice. A continuación se asigna a la base de datos original, la clasificación de las provincias en relación a cada clúster:

	province	poverty	gdp_per_cap	Cluster
1	Buenos Aires	8.167798	18.732051	3
2	Catamarca	9.234095	16.722352	3
3	Córdoba	5.382380	20.962931	2
4	Corrientes	12.747191	8.027456	5
5	Chaco	15.862619	9.317753	1
6	Chubut	8.051752	34.860686	4
7	Entre Ríos	7.288751	16.782775	3
8	Formosa	17.035583	7.180932	1
9	Jujuy	13.367965	9.631473	5
10	La Pampa	3.398774	21.916415	2
11	La Rioja	10.875152	16.756031	3
12	Mendoza	5.692798	19.225264	2
13	Misiones	13.529788	8.757160	5
14	Neuquén	9.456635	40.931431	4
15	Río Negro	8.678391	16.072442	3
16	Salta	16.870500	11.065861	1
17	San Juan	9.050784	12.131632	3
18	San Luis	6.593771	27.250930	2
19	Santa Cruz	8.024762	42.573981	4
20	Santa Fe	6.081012	25.540067	2
21	Santiago del Estero	11.759000	9.597026	5
22	Tucumán	11.214239	9.567956	5

Tabla 4.6. Valores del clúster agregados a la Tabla 1.1.

Con el objetivo de buscar un equilibrio entre la cohesión y separación de los grupos, se utiliza una medida de validación interna llamada *ancho de silueta media (si)* [34], la cual mide la distancia media con las observaciones del clúster vecino y la distancia media con las observaciones dentro de su propio clúster. Los valores próximos a cero indican mala agrupación, mientras que los cercanos a uno buena agrupación. Por otra parte, los valores negativos expresan una agrupación errónea y los valores por debajo del umbral de 0,25 representan una mala agrupación. El ancho de silueta, se representa en el próximo gráfico:

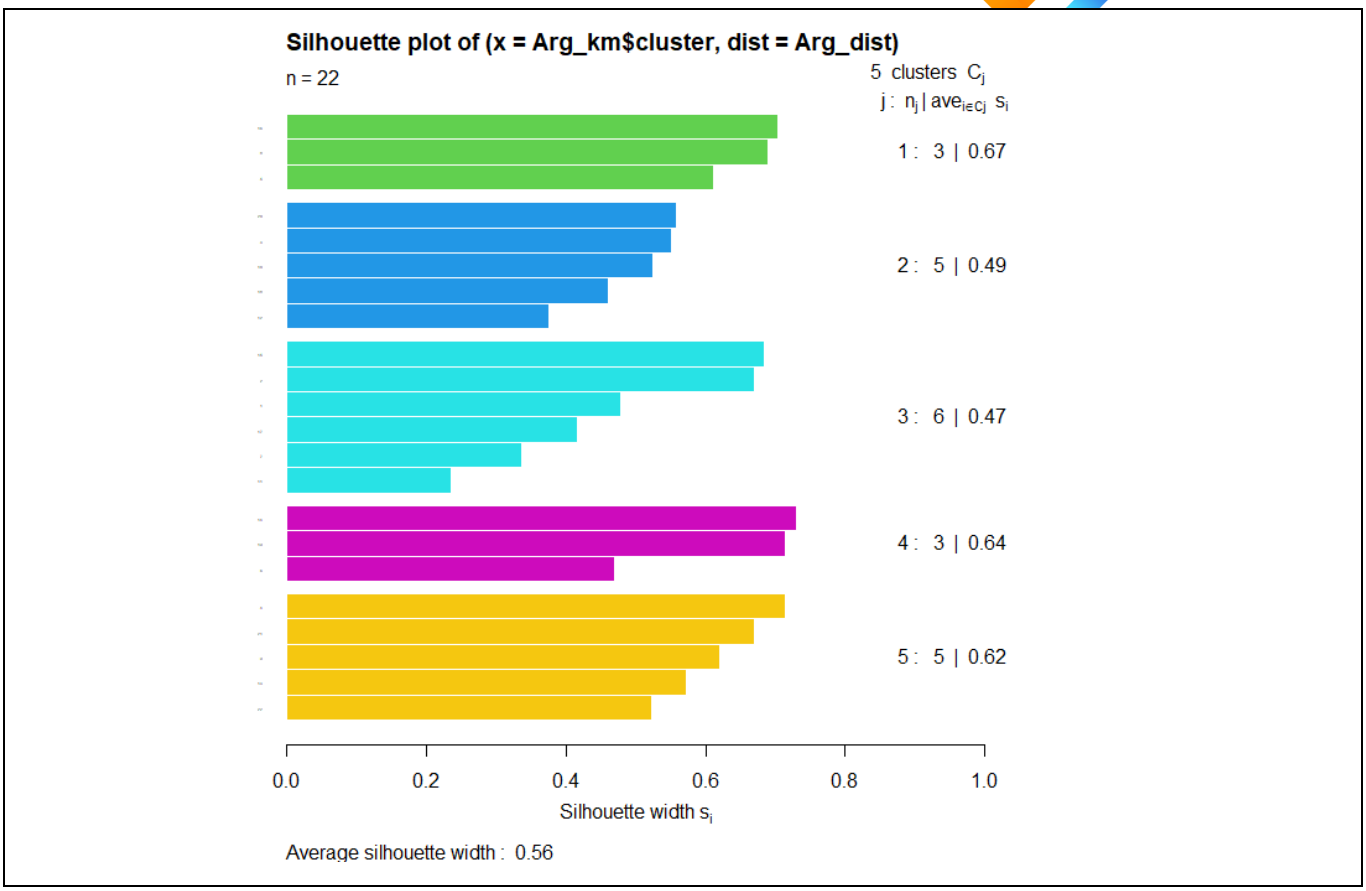


Figura 4.4. Validación interna del clúster.

Se observa en general que los grupos tienen una buena calidad de agrupación interna, cuyo valor medio es de 0.56. El clúster 1 está formado por 3 observaciones cuyo ancho de silueta es de 0.67. El grupo 2 y 3 son muy similares en cantidad de elementos, cuyos valores son 5 y 6 y con respecto a su validación interna sus valores varían entre 0.49 y 0.47. Por otra parte, el clúster 4 tiene 3 valores y un índice de 0.64, mientras que el quinto clúster presenta 5 elementos con un índice de silueta de 0.62. El clúster 1 presenta la mejor calidad de agrupación y el clúster 3 la más desfavorable.

Otra forma de visualizar los clúters es mediante una estructura de árbol llamada dendograma [28], el cual es un método de clustering jerárquico donde se muestra cierta jerarquía de agrupación como se puede notar a continuación:

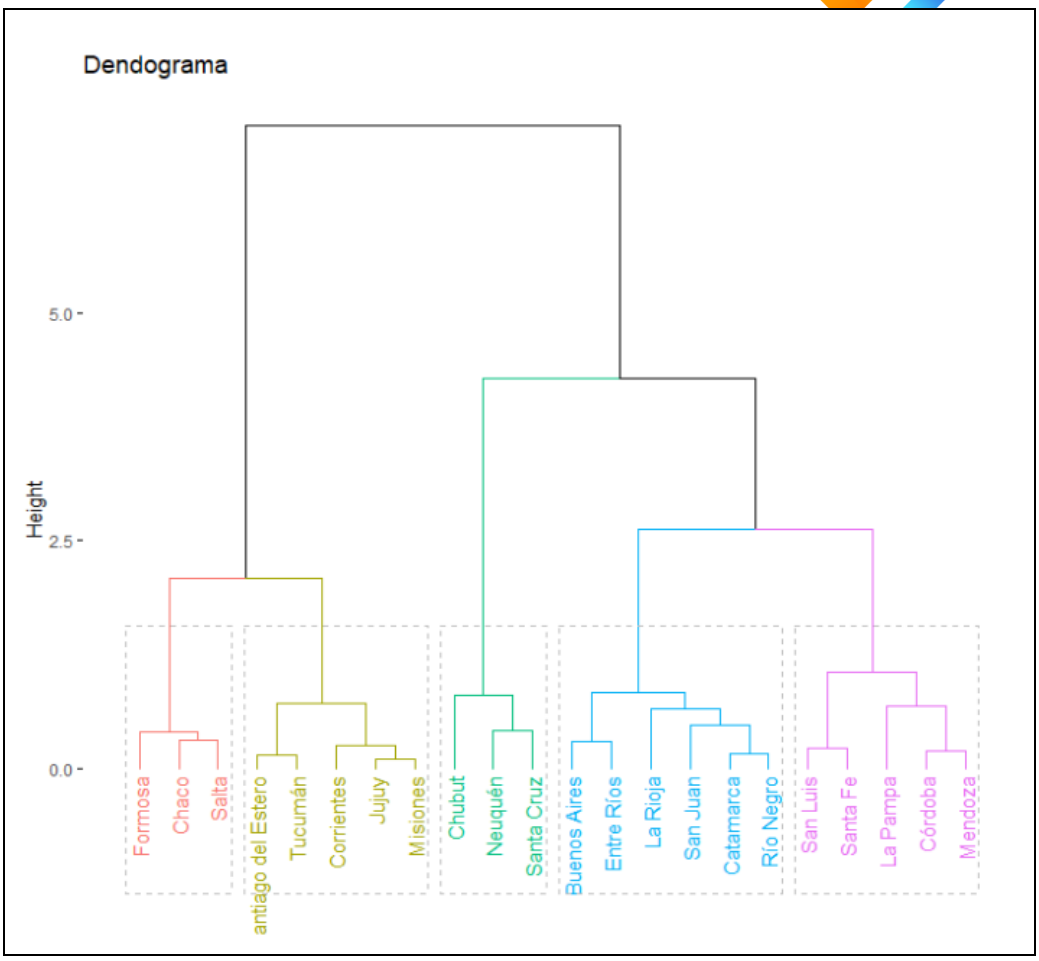


Figura 4.5. Visualización del clúster en dendogramas.

En el gráfico 4.5. se muestra cierta jerarquía de asociación entre los elementos, donde primero agrupa al clúster 1 y 2 y luego se divide entre el clúster 3, 4 y 5. El primer grupo está formado por: Formosa, Chaco y Salta; seguido de: Santiago del Estero, Tucumán, Corrientes, Jujuy y Misiones. En tercer lugar se encuentran: Chubut, Neuquén y Santa Cruz. En cuarto lugar: Buenos Aires, Entre Ríos, La Rioja, San Juan, Catamarca y Rio Negro. Finalmente se ubican: San Luis, Santa Fe, La Pampa, Córdoba y Mendoza.

También se puede mostrar el clustering jerárquico, mediante un gráfico similar a un árbol genealógico llamado filogénica.

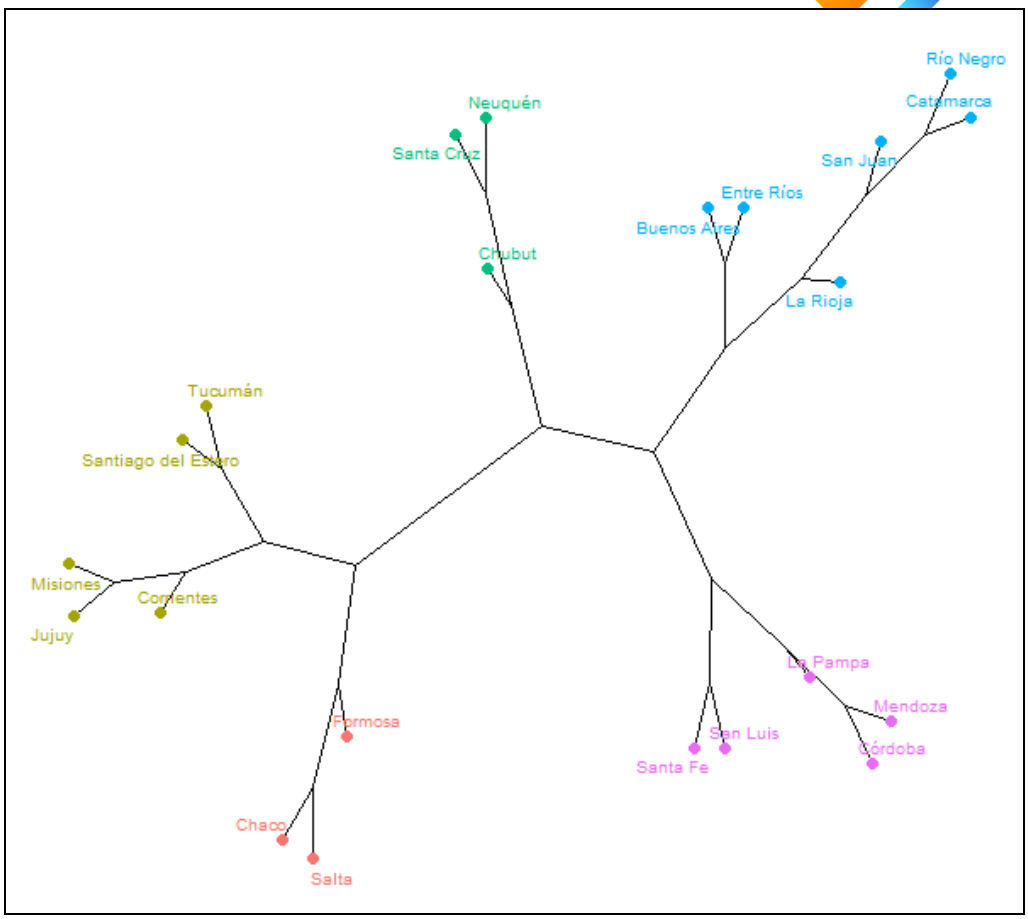


Figura 4.6. Representación filogénica del clúster.

A modo de resumen se observa que tanto el clúster jerárquico como no jerárquico permiten agrupar las provincias en relación al *PBI p/c* y al *índice de pobreza* en cinco clústers, siendo los clústers 1 y 5 los más desfavorables en términos de *PBI p/c* y de *nivel de pobreza* y los clústers 2, 3 y 4 los que poseen mejores indicadores en relación al *PBI p/c* y al bajo *índice de pobreza*.

Capítulo 5

ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales (PCA) se utiliza para reducir la dimensión de las magnitudes de un conjunto de datos minimizando la pérdida de información, de manera de representarlas a través de nuevas variables llamadas componentes principales [36] [37] [38], los cuales no deben estar correlacionados. Los componentes principales representan una combinación lineal de los factores, cuyos coeficientes representan los *autovalores* de la matriz de correlación, como se muestra a continuación:

$$Dim_j = a_{1j}.x_1 + a_{2j}.x_2 + \ldots + a_{pj}.x_k, \tag{7}$$

Donde Dim_j corresponde a cada componente principal y a_{pj} refleja los coeficientes de la combinación lineal entre las x_k variables observables. Los componentes principales, denotados por: $Dim_1, Dim_2, \ldots Dim_j$, se ordenan de mayor a menor en relación a la cantidad de varianza acumulada, la cual expresa la proporción de información capturada partir de las variables originales.

$$\sigma^2(Dim_1) \geq \sigma^2(Dim_2) \geq \cdots \geq \sigma^2\left(Dim_j\right), con\ j < k \tag{8}$$

En la Ecuación 8. se puede notar que la primera componente captura mayor proporción de varianza en relación a las componentes sucesivas [39] [40]. Para evitar fluctuaciones en el PCA, es importante escalar las variables en caso que existan magnitudes que provengan de diferentes unidades de medidas, para luego determinar la siguiente matriz de correlación entre las variables:

	gdp	illiteracy	poverty	deficient_infra	school_dropout	no_healthcare	birth_mortal	pop	movie_theatres_per_cap	doctors_per_cap	gdp_per_cap
gdp	1.00	-0.29	-0.25	-0.27	-0.18	-0.14	-0.11	0.99	-0.01	0.23	0.10
illiteracy	-0.29	1.00	0.64	0.58	0.62	0.79	0.34	-0.21	-0.63	-0.42	-0.65
poverty	-0.25	0.64	1.00	0.50	0.38	0.69	0.47	-0.16	-0.65	-0.64	-0.58
deficient_infra	-0.27	0.58	0.50	1.00	0.19	0.63	0.38	-0.23	-0.27	-0.35	-0.38
school_dropout	-0.18	0.62	0.38	0.19	1.00	0.42	0.42	-0.16	-0.44	-0.18	-0.42
no_healthcare	-0.14	0.79	0.69	0.63	0.42	1.00	0.35	-0.05	-0.68	-0.43	-0.75
birth_mortal	-0.11	0.34	0.47	0.38	0.42	0.35	1.00	-0.08	-0.04	-0.19	-0.34
pop	0.99	-0.21	-0.16	-0.23	-0.16	-0.05	-0.08	1.00	-0.08	0.16	-0.01
movie_theatres_per_cap	-0.01	-0.63	-0.65	-0.27	-0.44	-0.68	-0.04	-0.08	1.00	0.40	0.60
doctors_per_cap	0.23	-0.42	-0.64	-0.35	-0.18	-0.43	-0.19	0.16	0.40	1.00	0.30
gdp_per_cap	0.10	-0.65	-0.58	-0.38	-0.42	-0.75	-0.34	-0.01	0.60	0.30	1.00

Tabla 5.1. Matriz de correlación de variables estandarizadas

En la Tabla 5.1. para poder identificar el grado de asociación entre las variables numéricas de la base de datos, se aplica el coeficiente de correlación de Pearson [7] [8], cuyo indicador varía en un rango entre -1 y 1. Los valores próximos a -1 indican correlación negativa fuerte, mientras que aquellos próximos a +1 reflejan correlación positiva fuerte. En general se observa que el grado de asociación entre las variables es positiva media y en menor proporción negativa débil. Dado que los coeficientes son distintos a cero, se comprueba que las variables están correlacionadas por lo tanto se puede aplicar el modelo (PCA).

Para visualizar como se correlacionan las variables, se muestra el siguiente gráfico de correlación:

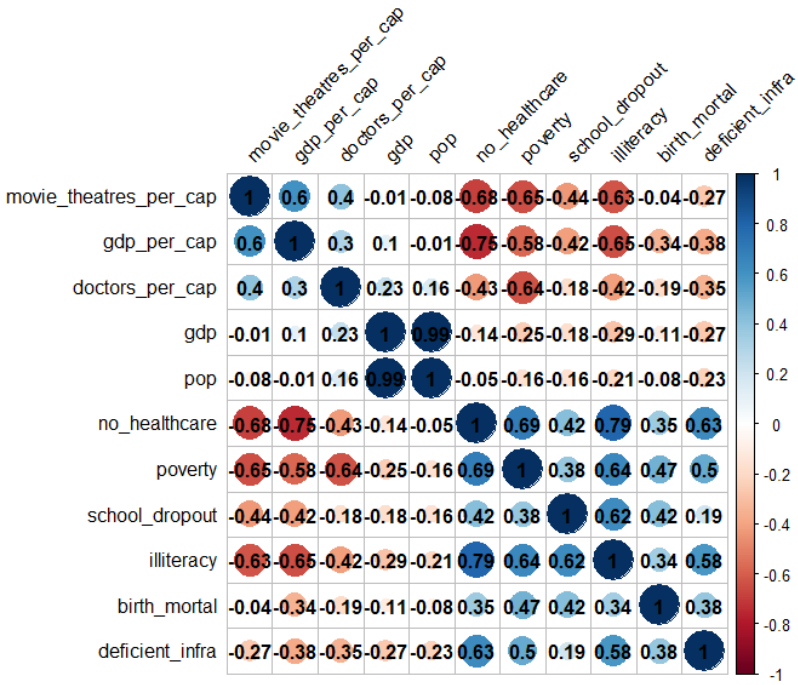


Figura 5.1. Correlograma.

En la Figura 5.1. el color azul representa correlación positiva, mientras que el color rojo refleja correlación negativa, cuya intensidad indica correlación fuerte y los colores más tenues muestran correlación débil. Por ejemplo, el *PBI p/c* presenta una fuerte correlación positiva con la *cantidad de películas p/c*, cuyo valor es de 0.60, mientras que su grado de asociación con respecto a la *falta de acceso a la salud* es de -0.75, cuya correlación es inversa y fuerte.

5.1. Supuestos del modelo (PCA).

Además, se deben verificar los siguientes supuestos: prueba de esfericidad de Bartlett [41], prueba de adecuación muestral (KMO) [42] [43] y la prueba del determinante. La prueba de esfericidad de Bartlett, evalúa la naturaleza de las correlaciones y parte de las siguientes hipótesis:

Ho: No existe colinealidad entre las variables, en cuyo caso no se podría aplicar (PCA).

H1: Hay colinealidad entre las variables.

Al implementar la prueba de Bartlett, el p-valor es de 2.685798e-18. Como el resultado es menor al 5%, se rechaza la Ho, por lo tanto las variables son colineales. Por otra parte, la prueba de adecuación muestral de Kaiser-Meyer-Olkim (KMO), evalúa si existe una buena adecuación de los datos al (PCA). Su valor oscila entre 0 y 1. Los valores próximos a 0 indican mala adecuación, mientras que los valores cercanos a 1 reflejan buena adecuación.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(Arg_scale))
Overall MSA =  0.59
MSA for each item =
                   gdp             illiteracy                poverty        deficient_infra
                  0.33                   0.68                   0.78                   0.56
        school_dropout          no_healthcare           birth_mortal                    pop
                  0.46                   0.90                   0.57                   0.31
movie_theatres_per_cap        doctors_per_cap           gdp_per_cap
                  0.72                   0.64                   0.56
```

Tabla 5.2. Prueba Kaiser-Meyer-Olkin.

El valor global de la prueba es de 0.59=59%, cuya adecuación muestral es regular. Por otra parte, la adecuación parcial de cada variable en general es regular, con excepción de las variables *PBI* y *población*, cuyos valores son: 0.33 y 0.31 siendo muy poco representativos para el análisis de PCA.

Por último se aplica la prueba del determinante para evaluar si existe multicolinealidad entre las variables. El valor del determinante de la matriz de correlación es de 5.538618e-06. Como su valor es positivo y próximo a cero se verifica la prueba del determinante.

En vista a que se corroboran las pruebas anteriores se aplica a continuación el método de componentes principales:

```
Call:
PCA(X = Arg_scale, scale.unit = TRUE, ncp = 2, graph = TRUE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9
Variance               4.864   1.062   0.987   0.743   0.533   0.334   0.185   0.160   0.132
% of var.             54.039  11.799  10.971   8.258   5.917   3.715   2.053   1.780   1.468
Cumulative % of var.  54.039  65.838  76.809  85.067  90.984  94.699  96.752  98.532 100.000
```

Tabla 5.3. Método de PCA.

En la Tabla 5.3. se observa que la primera componente captura 54.039% de varianza explicada, seguida de la segunda y tercer componente, cuyos porcentajes de varianza son : 11.799% y 10.971% respectivamente. Para seleccionar la cantidad de componentes principales a retener, se puede utilizar el método de Keiser-Guttman [39] [44], un gráfico de sedimentación o el análisis de Horn [45].

El método de Keiser-Guttman determina que los componentes principales son significativos si los autovalores son mayores a uno y si el porcentaje de varianza acumulada es mayor al 60%. Los resultados de la prueba, son:

```
       eigenvalue variance.percent cumulative.variance.percent
Dim.1  4.8635391        54.039323                    54.03932
Dim.2  1.0618654        11.798505                    65.83783
Dim.3  0.9874007        10.971119                    76.80895
Dim.4  0.7432071         8.257857                    85.06680
Dim.5  0.5325174         5.916860                    90.98366
Dim.6  0.3343400         3.714889                    94.69855
Dim.7  0.1847902         2.053224                    96.75178
Dim.8  0.1602237         1.780263                    98.53204
Dim.9  0.1321164         1.467960                   100.00000
```

Tabla 5.4. Método de Keiser-Guttman.

En la Tabla 5.4. se comprueba que los autovalores son mayores a uno para las dos primeras componentes, cuyo porcentaje de varianza acumula es de 65.83783%, el cual es superior al 60% propuesto por la prueba. Por lo tanto se podrían elegir los componentes principales: *Dim1*, *Dim2*. Otra forma de determinar la cantidad de componentes principales es mediante el siguiente gráfico de sedimentación:

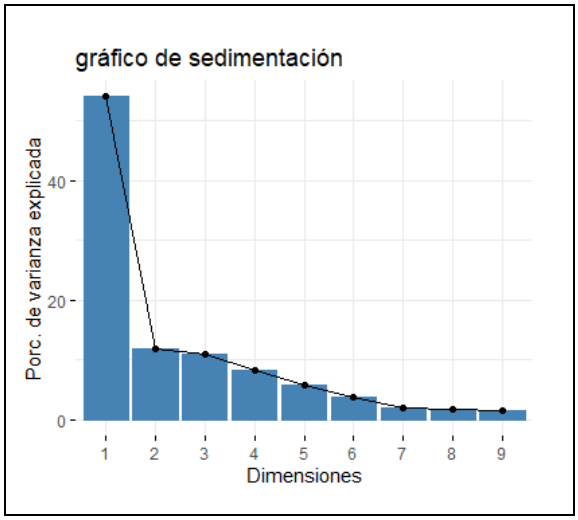


Figura. 5.2. Gráfico de sedimentación.

En la Figura 5.2. se puede notar que los componentes a la izquierda del punto de inflexión son significativos. Además, se observa que a partir del segundo componente principal se produce un quiebre de la curva llamado codo. Por lo tanto se corrobora la elección de dos componentes. Por último se aplica el análisis de Horn, mediante la siguiente representación:

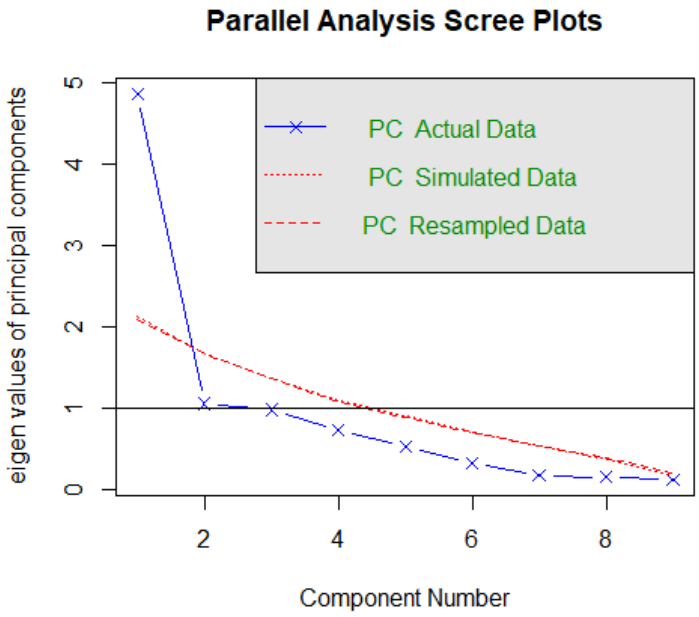


Figura. 5.3. Análisis de Horn.

Al observar la intersección entre la línea roja, que representa los valores por azar y la línea azul los valores actuales se obtienen 2 componentes principales. De acuerdo a los métodos analizados se utilizarán 2 componentes principales al método (PCA).

5.2. Características de cada variable con los componentes principales.

En el próximo apartado se explican los principales fundamentos presentes en análisis de componentes principales: la correlación, la calidad de representación, la contribución porcentual de cada variable a los componentes principales.

1. **Correlación de cada variable con los componentes principales.**

El grado de asociación entre cada variable y los componentes principales, se representa mediante la siguiente tabla:

	Dim.1	Dim.2
illiteracy	0.8811956	-0.01720908
poverty	0.8523898	-0.03931511
deficient_infra	0.6470718	0.25155076
school_dropout	0.6047098	0.25016043
no_healthcare	0.8990589	-0.08520575
birth_mortal	0.4990576	0.78495585
movie_theatres_per_cap	-0.7466152	0.50052401
doctors_per_cap	-0.5873966	0.22680458
gdp_per_cap	-0.7850189	0.09372982

Tabla 5.5. Matriz de correlación entre variables y componentes principales.

En base a la Tabla 5.5. se puede observar por ejemplo que la variable *PBI p/c* denotada como *"gdp_per_cap"* se correlaciona de manera negativa y fuerte con la componente 1 , mientras que con las componentes 2 se relaciona de manera positiva y débil.

2. **Calidad de la representación.**

La calidad de la representación de cada magnitud en relación a los componentes principales, se muestra en la próxima tabla:

	Dim.1	Dim.2
illiteracy	0.7765057	0.0002961525
poverty	0.7265684	0.0015456779
deficient_infra	0.4187019	0.0632777845
school_dropout	0.3656740	0.0625802399
no_healthcare	0.8083070	0.0072600205
birth_mortal	0.2490585	0.6161556831
movie_theatres_per_cap	0.5574342	0.2505242814
doctors_per_cap	0.3450347	0.0514403197
gdp_per_cap	0.6162547	0.0087852792

Tabla 5.6. Calidad de representación entre las variables y componentes principales.

Se observa que las variables están mejor representadas por la componente Dim.1 que por la componente Dim.2.

3. Contribución porcentual.

A fin de evaluar la proporción que cada magnitud aporta a los componentes principales, se muestra la siguiente matriz:

	Dim.1	Dim.2
illiteracy	15.965857	0.02788983
poverty	14.939087	0.14556250
deficient_infra	8.608996	5.95911517
school_dropout	7.518681	5.89342469
no_healthcare	16.619728	0.68370438
birth_mortal	5.120932	58.02577810
movie_theatres_per_cap	11.461493	23.59284635
doctors_per_cap	7.094314	4.84433506
gdp_per_cap	12.670912	0.82734393

Tabla 5.7. Contribución de cada variable a los componentes principales.

En la Tabla 5.7. la variable *PBI p/c* y *pobreza* están mejor representadas por la primera componente, mientras que las magnitudes: *mortalidad infantil* y *cantidad de películas p/c* son mejor caracterizadas por la segunda componente. Para poder visualizar que variables tienen mayor aporte a cada componente, se muestran los siguientes gráficos:

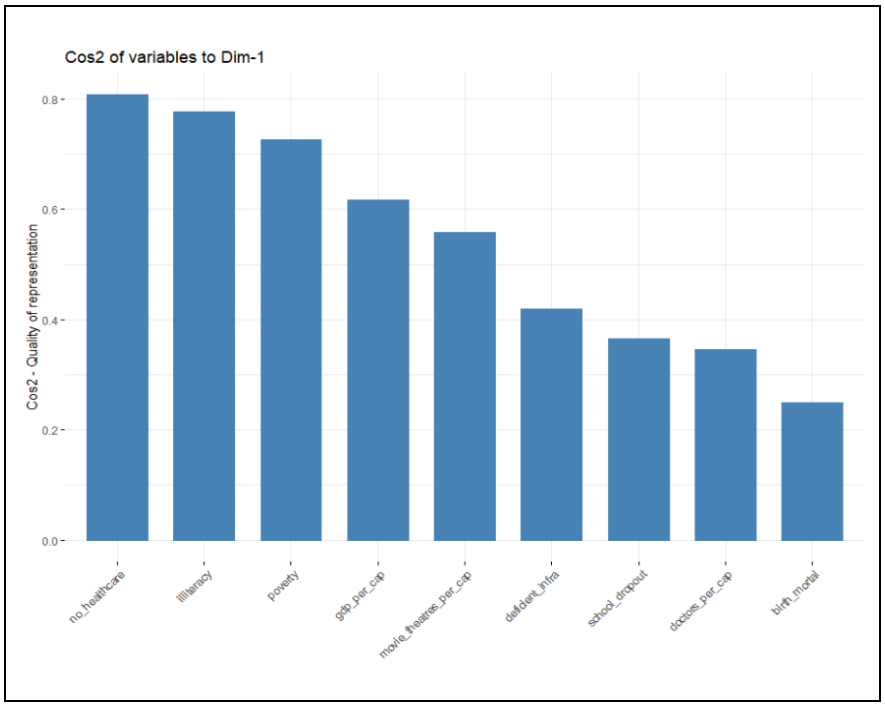


Figura. 5.4. Cos2 de la componente 1.

Las variables con mayor calidad de representación en la primera componente son*: falta de acceso a la salud, analfabetismo, pobreza* y *PBI p/c.*

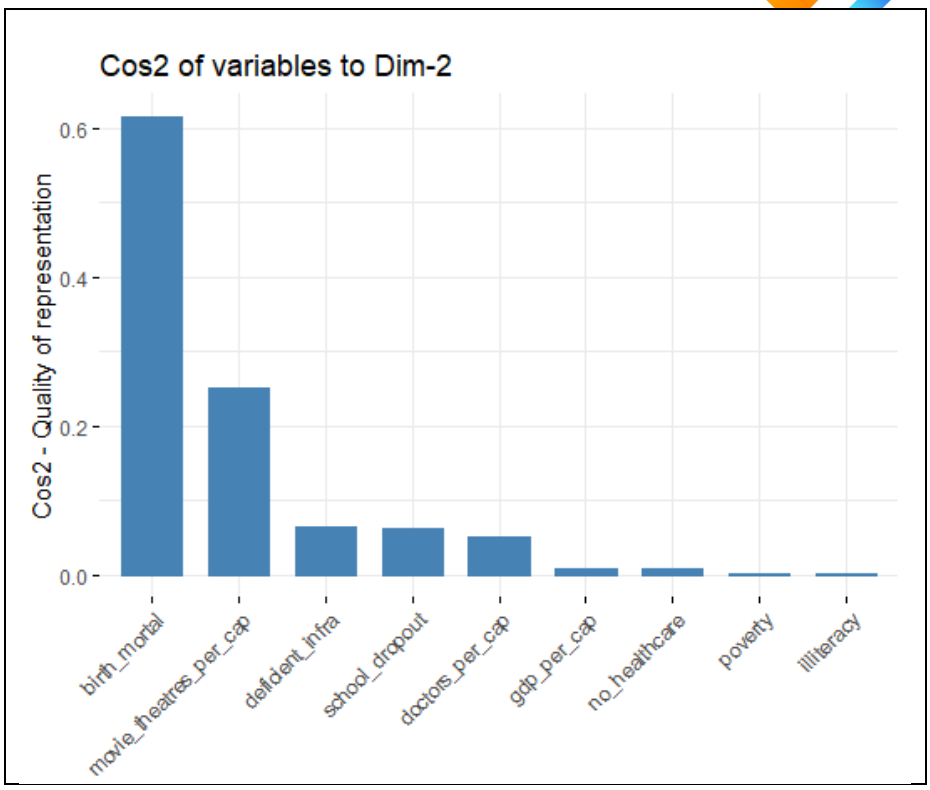


Figura. 5.5. Cos2 de la componente 2.

En la segunda componente, las magnitudes que poseen una calidad de representación son: *mortalidad infantil* y *cantidad de películas p/c*. A continuación se presentan las variables más significativas para cada componente principal:

```
$Dim.1

Link between the variable and the continuous variables (R-square)
================================================================================
                      correlation      p.value
no_healthcare           0.8990589 1.299274e-08
illiteracy              0.8811956 6.138581e-08
poverty                 0.8523898 4.749925e-07
deficient_infra         0.6470718 1.134395e-03
school_dropout          0.6047098 2.870888e-03
birth_mortal            0.4990576 1.805731e-02
doctors_per_cap        -0.5873966 4.048315e-03
movie_theatres_per_cap -0.7466152 6.578434e-05
gdp_per_cap            -0.7850189 1.513423e-05
```

```
$Dim.2

Link between the variable and the continuous variables (R-square)
================================================================================
                      correlation      p.value
birth_mortal            0.7849558 0.0000151744
movie_theatres_per_cap  0.5005240 0.0176683448
```

Tabla 5.8. Variables significativas para los dos primeros componentes principales.

Análisis de individuos

Luego de identificar que variables tienen mayor preponderancia para cada componente, es importante identificar que grupos de individuos poseen características comunes. Para ello se muestra el siguiente gráfico:

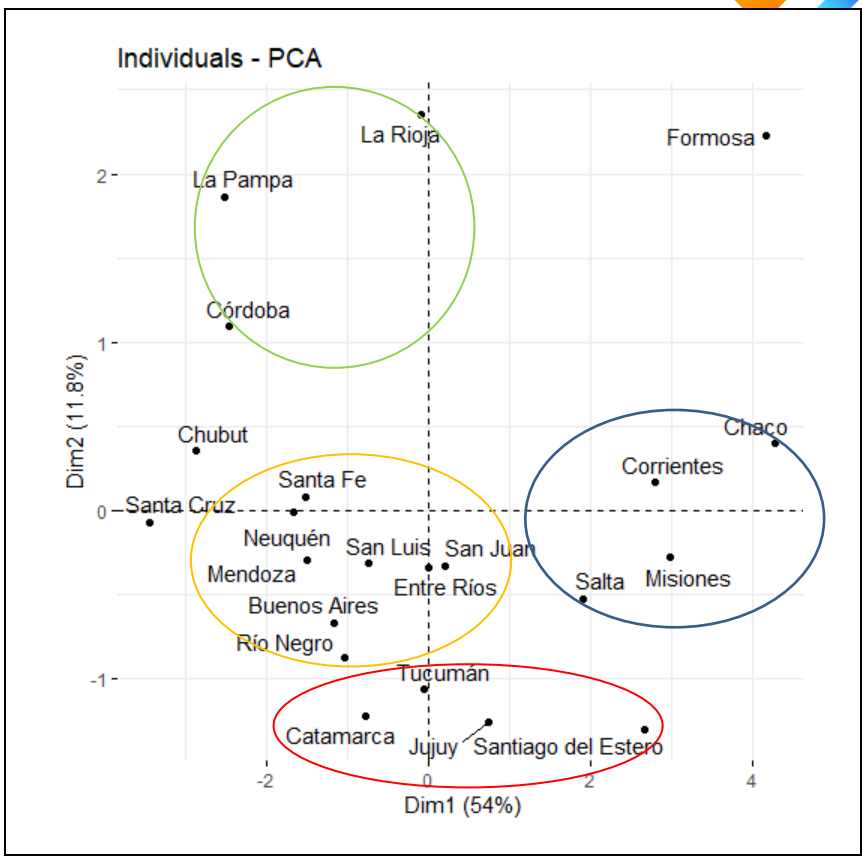


Figura. 5.6. PCA. Análisis de individuos.

En la Figura 5.6. se observan las provincias que presentan características socioeconómicas semejantes dado que sus puntos están muy próximos entre sí. El primer grupo está formado por: Corrientes, Chaco, Misiones, Salta, luego el segundo grupo formado por: Santa Fe, Neuquén, Mendoza, Buenos Aires Entre Ríos, San juan, San Luis, seguido por las provincias: Tucumán, Catamarca, Jujuy y Santiago del Estero y finalmente las provincias de: Córdoba, La Rioja y La Pampa. Luego, para mostrar la calidad de representación de cada individuo con respecto a los componentes principales, se presentan los siguientes gráficos:

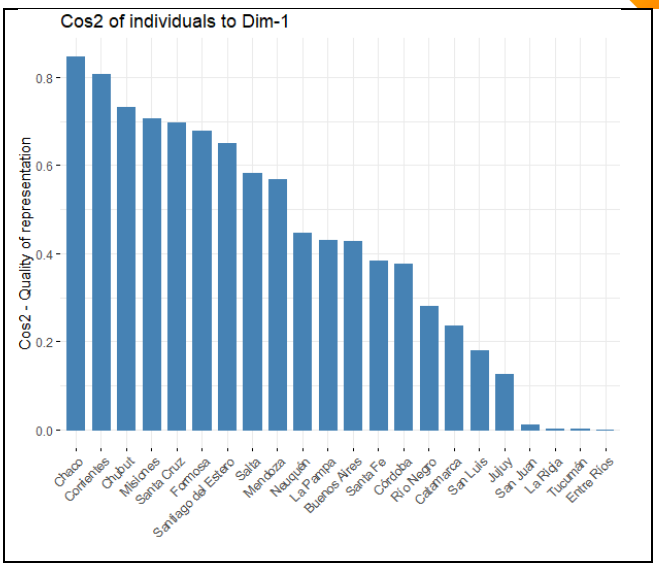


Figura 5.7. Calidad de representación de individuos en el primer componente.

En la Figura 5.7. se observan que los elementos que superan un 50% de la calidad de representación con respecto al primer componente principal son: Chaco, Corrientes, Chubut, Misiones, santa cruz, Formosa, Santiago del Estero, Salta y Mendoza.

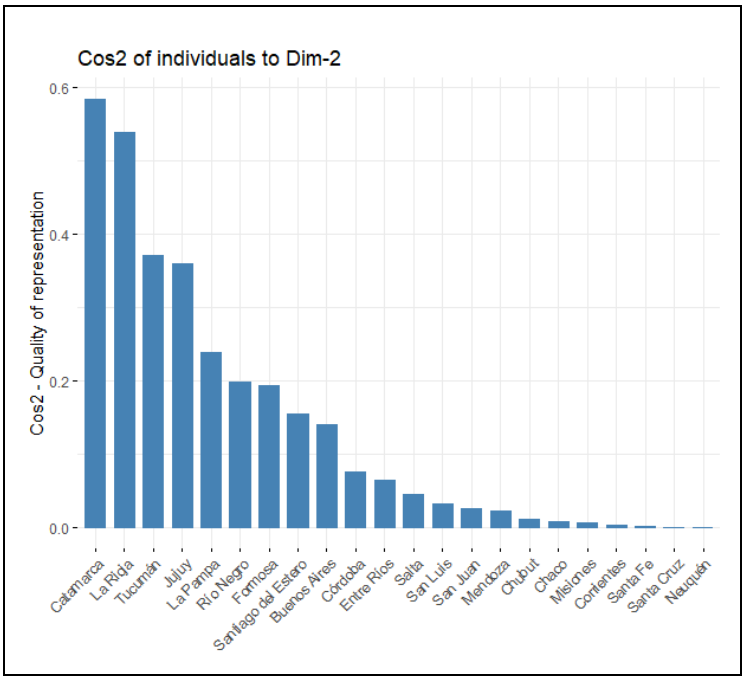


Figura 5.8. Calidad de representación de individuos en el segundo componente.

Las provincias con mayor calidad de representación de acuerdo a la componente principal 2 son: Catamarca, La Rioja, Tucumán y Jujuy, cuya proporción de representación es de aproximadamente de un 50%.

Análisis de las variables

A continuación, se evalúa en nivel de correlación entre las variables en función a cada componente principal, a través del siguiente gráfico:

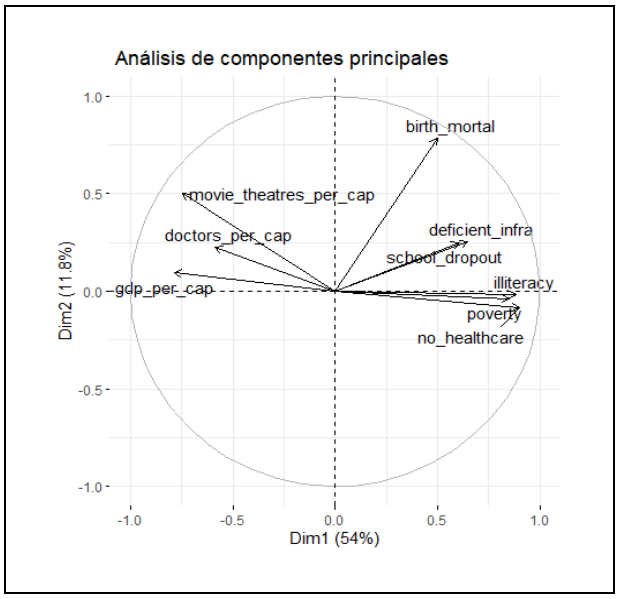


Figura. 5.9. Cirulo de correlación de las variables

En base al círculo de correlación, se observa que el *PBI p/c, cantidad de médicos p/c* y *películas p/c* presentan correlación positiva y fuerte, mientras que el *PBI p/c* y *el analfabetismo* poseen un nivel de asociación negativa fuerte. Por otra parte, la *mortalidad infantil* en relación a la *cantidad de médicos* y *de películas per cápita* casi no presentan asociación. Finalmente, para relacionar el comportamiento entre individuos y las variables, se muestra el siguiente gráfico:

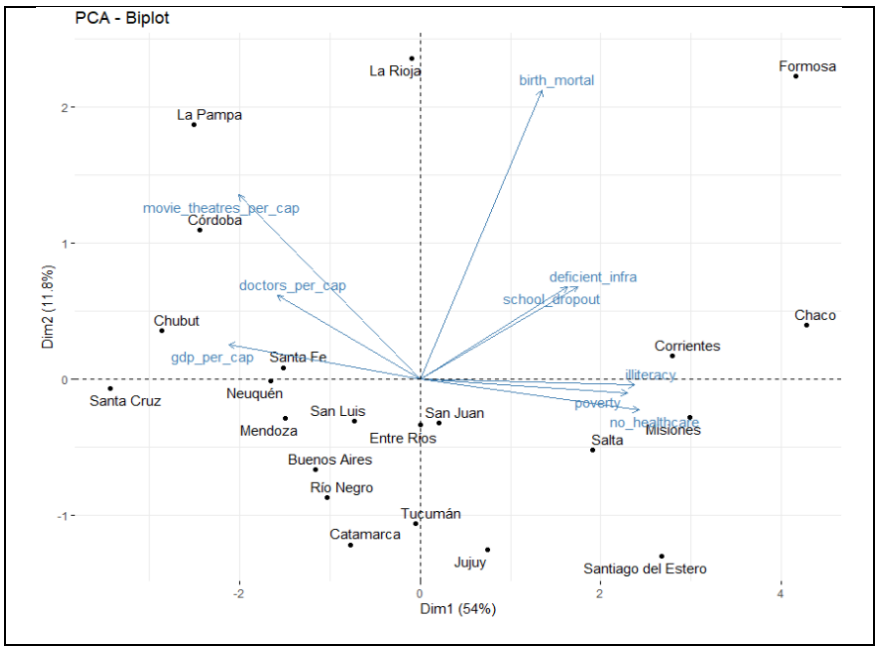


Figura. 5.10. Gráfico de Biplot.

En la Figura 5.10. las provincias de: Córdoba y La Pampa, son más representativas con respecto a la *cantidad de médicos* y *de películas per cápita*. Por otra parte, el *PBI p/c* está más caracterizado por la provincia de: Chubut, Neuquén y Santa Fe y es menor en las provincias de: Corrientes, Misiones y Salta, donde prevalecen altos índices de: *analfabetismo, falta de acceso a la salud* y *pobreza*.

También, se puede filtrar una lista con las cinco variables más importantes para el modelo PCA, como se muestra a continuación:

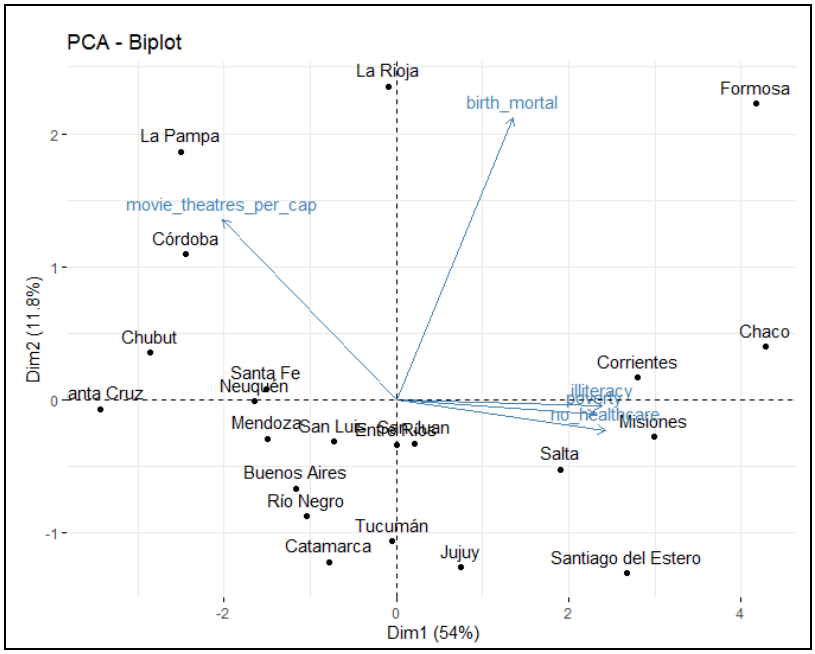


Figura. 5.11. Variables más importantes del PCA.

En base a la Figura 5.11. las magnitudes que mejor representan el comportamiento del primer componente principal son: *pobreza, falta de acceso a la salud* y *analfabetismo* presentes en Salta, Misiones y Corrientes. Corrientes, Misiones y Salta. Por otra parte, la *mortalidad infantil* está caracterizada por la provincia de La Rioja y la *cantidad de películas p/c* correspondiente a: Córdoba y La Pampa.

RESULTADOS

En base al correlograma de la Figura 5.1. se observa que el *PBI p/c* tiene una correlación fuerte y positiva con respecto a la *cantidad de películas p/c* y positiva débil en relación a la *cantidad de médicos p/c*. Por otra parte, presenta un grado de asociación negativo fuerte frente a la *falta de acceso a la salud, a la pobreza* y *al analfabetismo*, mientras que su correlación es negativa débil en relación a la *mortalidad infantil* y *deficiencia de infraestructura*.

Luego, al implementar una prueba paramétrica para evaluar la influencia del *PBI p/c* en cada región, se tiene que la media del *PBI p/c* presenta diferencias significativas. De acuerdo a la Figura 2.2. La menor diferencia se observa entre las regiones centro y norte, mientras que la mayor variación se da entre los grupos norte y sur. Por otra parte, al aplicar un modelo de regresión lineal múltiple, de acuerdo a la prueba "Stepwise", los predictores óptimos para la variable respuesta *PBI p/c* son: el *PBI, deserción escolar, población, falta de acceso a la salud* y *médicos p/c*. Para un nivel de confianza del 95%, se tiene que el R^2 es de 0.735 por lo tanto el modelo de regresión lineal múltiple es estadísticamente significativo. Entonces, al evaluar el R^2parcial de los predictores, de acuerdo a la Tabla 3.3. se puede notar que las variables más influyentes en el modelo son la *población* y el *PBI* y en menor medida: la *deserción escolar, médicos per cápita* y *falta de acceso a la salud*.

Además, se verifican los supuestos de linealidad según la figura 3.1, y de normalidad mediante el grafico cuantil-cuantil perteneciente a la Figura 3.2. Por otra parte, no se no se verifica la homocedasticidad de acuerdo a la Figura 3.2.

A continuación, se utiliza un modelo de regresión logística con el fin de hallar la probabilidad que una provincia presente un valor de *PBI p/c* superior a su media, se observa que cuando el valor del *PBI p/c* de una provincia aumenta, se incrementa la probabilidad que su valor sea superior a la media de acuerdo a la Tabla 3.6. Además en base a la Figura 3.3. se observa que la distribución de los residuos es normal, no obstante se comprueba la presencia de valores atípicos.

Posteriormente, se genera un modelo de clustering no jerárquico y se obtienen 5 grupos, formados por 3, 5, 6, 3 y 5 elementos respectivamente, siendo la calidad de agrupación muy buena correspondiente a 94.3%. El primer y el quinto clúster poseen mayor índice de pobreza, mientras que los clústers 2,3 y 4 tienen valores medios y alto de *PBI p/c* y niveles medios de *pobreza*. Al evaluar la calidad de agrupación interna mediante el grafico de silueta presente en la Figura 4.4. se observa que los grupos presentan una buena calidad de agrupación interna promedio correspondiente a 0.56. El clúster 1 presenta el mayor valor del ancho de silueta,

mientras que el clúster 3 tiene el menor, cuyos índices son 0.67 y 0.47 respectivamente. Por otro lado, al crear un clúster jerárquico, no se observan resultados diferentes en la representación de los grupos según el dendograma de la Figura. 4.5.

Por último, se implementa un modelo para reducir la dimensión de los predictores y se obtienen dos componentes principales, cuya calidad de representación de las variables es mayor en la primera componente y menor en la segunda. Además, en base a la Tabla 5.5. se observa que el *PBI p/c* se correlaciona de manera negativa y fuerte con la componente 1 mientras que con las componentes 2 se relaciona de manera positiva y débil. Por otra parte, en base al círculo de correlación presente en la Figura 5.9. se puede notar que el *PBI p/c*, *cantidad de médicos p/c y películas p/c* presentan correlación positiva y fuerte, mientras que el *PBI p/c* y el *analfabetismo* tienen un correlación negativa fuerte.

En resumen, de acuerdo a la Figura 5.10. se observa que el *PBI p/c* es más representativo en las provincias de: Chubut, Neuquén y Santa Fe y es menor en las provincias de: Corrientes, Misiones y Formosa, donde además prevalecen altos índices de: *analfabetismo, falta de acceso a la salud* y *pobreza*.

DISCUSIÓN

Considerando las limitaciones con respecto a la cantidad de datos presentes en este trabajo, se observa que el *PBI p/c* se relaciona de manera fuerte y directa con respecto a la *cantidad de películas p/c* y directa débil con la *cantidad de médicos p/c*, es decir, al aumentar el *PBI p/c* por provincia aumenta considerablemente la *cantidad de películas p/c* y se incrementa en menor medida con respecto a la *cantidad de médicos per cápita*. Por otra parte, su grado de asociación es inverso con el resto de las magnitudes, con excepción de la *población* y del *PBI*. Por ejemplo, al incrementar los indicadores de: *pobreza, deserción escolar y analfabetismo*, disminuye el *PBI p/c*. Bajo un punto de vista económico, se comprueba que el *PBI per cápita* promedio en la región sur es mayor que en la región centro y norte del país, debido a un mayor nivel de desarrollo productivo presente en los sectores primarios y secundarios.

Al evaluar el efecto del *PBI p/c* y el *índice de pobreza* en las provincias, se observa que la región norteña tiene mayor *índice de pobreza* y un bajo nivel de *PBI p/c*, como es el caso de: Formosa, Salta y Chaco, donde además prevalecen valores considerables de *analfabetismo* y *falta acceso sanitario*. Por otra parte, las provincias ubicadas en las regiones centro y sur presentan mejores indicadores de *PBI p/c*, mayor *cantidad de médicos per cápita* y *menores índices de pobreza*, como se observa en: La Pampa, Córdoba, Santa Fe, Chubut, Santa Cruz y Neuquén.

Estos hallazgos exponen implicancias no solo económicas, sino también políticas y sociales reflejando menor nivel de desarrollo económico, mayor diferencia social y falta de inversión en la región norte del país. En contrapartida, la región centro posee un dinamismo económico mayor pero con niveles socioeconómicos medios. Por último, la región sur del país es la más favorecida en materia de desarrollo, inversión y estándar de vida dado que posee mejores indicadores tanto económicos como sociales.

CONCLUSIONES

En este trabajo de investigación se han analizado los factores influyentes en el producto bruto interno per cápita de diferentes regiones y provincias de la Argentina. Se ha podido constar que la región norte del país es la más desfavorable en materia de desarrollo socioeconómico, donde prevalece un bajo índice de *PBI per cápita*, altos indicadores de *pobreza, analfabetismo* y *falta de acceso al servicio sanitario*. Por el contrario, las regiones centro y sur mostraron mejor desempeño en relación al *PBI per cápita* asociado a un menor *índice de pobreza*, mejor *acceso al sistema de salud* y mayor *cantidad de médicos por habitante*.

BIBLIOGRAFÍA

[1] González, L. (2009). ***Vulnerabilidad social y dinámica demográfica en Argentina.*** 2001-07. https://www.researchgate.net/publication/43071340_VULNERABILIDAD_SOCIAL_Y_DINAMICA_DEMOGRAFICA_EN_ARGENTINA_2001-07

[2] CEDLAS (2003). ***Boletín Semestral de Estadísticas Distributivas***. Centro de Estudios Distributivos, Laborales y Sociales. Departamento de Economía de la Universidad Nacional de La Plata.

[3] Cuaderno de Economía 55 (2001). ***Características Regionales y Sectoriales del Empleo y del Desempleo***. Ministerio de Economía de la Provincia de Buenos Aires.

[4] Triola, Mario. (2009). ***Estadística***. Pearson Educación. Cap.10, 11, 12.

[5] Spiegel, M. R. (1970). ***Estadística.*** Mc Graw-Hill. Cap.10, 13, 14,15.

[6] Lind, D. A. Et al. (2012). ***Estadística aplicada a los negocios y a la economía***. Mc Graw-Hill. Cap.12, 13, 14.

[7] Devore, J. L. (2018). ***Fundamentos de probabilidad y estadística***. Cengage. Vol.1, Cap. 7, 8, 9.

[8] Da Costa Pereira, N. Et al. (2006). ***La matriz de correlación***: ***una dicotomía entre soporte estadístico y herramienta agenciada.*** I Encuentro Latinoamericano de metodología en Ciencias Sociales. Mesa E-2.- Estado actual de los métodos/técnicas cuantitativas y cualitativas y de la triangulación metodológica. Ponencia: Universidad Nacional de Lujan. Buenos Aires, Argentina. http://sedici.unlp.edu.ar/handle/10915/109909

[9] Garibaldi, L. A. Et al. (2019). ***Modelos estadísticos en lenguaje R.*** Editorial Universidad Nacional de Rio Negro. http://rid.unrn.edu.ar/handle/20.500.12049/5789

[10] Timm, N. H. (2002). ***Applied multivariate analysis***. Cap.3. https://doi.org/10.1007/978-0-387-22771-9_4

[11] Gómez-Gómez, M., Danglot-Banck, C., & Vega-Franco, L. (2003). ***Sinopsis de pruebas estadísticas no paramétricas. Cuándo usarlas***. *Revista mexicana de pediatría*, *70*(2). https://www.medigraphic.com/cgibin/new/resumen.cgi?IDARTICULO=8084

[12] Ramalle-Gómara, E., & De Llano, J. A. (2003). ***Utilización de métodos robustos en la estadística inferencial.*** Atención primaria, 32(3), 177. https://core.ac.uk/download/pdf/82434100.pdf

[13] Hanusz, Z. Et al. (2016). ***Shapiro–Wilk Test with known mean***. REVSTAT-statistical journal. Vol.14, No.1, pp. 89–100. https://doi.org/10.57805/revstat.v14i1.180

[14] Brunetzarza, K. (2019). ***Distribución normal y algunas aplicaciones***. Universidad Autónoma del Estado de México. http://ri.uaemex.mx/handle/20.500.11799/106113

[15] Wilcox, R. (2012). ***Introduction to Robust Estimation and Hypothesis Testing*** (3rd ed.). Elsevier.

[16] Patil, I. (2021). ***Visualizations with statistical details: The 'ggstatsplot' approach***. Journal of Open Source Software, 6(61), 3167. https://doi:10.21105/joss.03167

[17] Dagnino, J. S. (2014). ***Regresión lineal.*** Revista chilena de anestesia. Vol.43, Nro.2. pp.143-149. https://doi.org/10.25237/revchilanestv43n02.14

[18] Astorga Gómez, J. M. (2014). ***Aplicación de modelos de regresión lineal para determinar las armónicas de tensión y corriente***. Ingeniería Energética. Vol.35, No.3, pp.234-241. https://www.redalyc.org/pdf/3291/329132445008.pdf

[19] Ramajo Hernández, J. (1994). ***Estimación de respuestas renta de los consumidores: una aplicación del modelo de regresión lineal discontinua a tramos***. Estudios de Economía Aplicada. Vol.1, No.1, pp.61-86. https://dialnet.unirioja.es/servlet/articulo?codigo=175950

[20] Chan, D. Et al. (2019). ***Análisis inteligente de datos con lenguaje R: con aplicaciones e imágenes.*** Ciudad de Buenos Aires.UTN-edUTecNe. https://ria.utn.edu.ar/xmlui/handle/20.500.12272/4371?show=full

[21] Rigalli, A. Et al. (2019). ***Uso de herramientas informáticas para la recopilación, análisis e interpretación de datos de interés en las ciencias biomédicas: estadística básica con R***. Facultad de Ciencias médicas. Universidad Nacional de Rosario. No.1, pp 76-105. https://rephip.unr.edu.ar/bitstream/handle/2133/15296/libroRmodulo3.pdf?sequence=3&isAllowed=y

[22] Garibaldi, L. A. Et al. (2019). ***Modelos estadísticos en lenguaje R.*** Editorial Universidad Nacional de Rio Negro. http://rid.unrn.edu.ar/handle/20.500.12049/5789

[23] Piol, R. (2014). ***Validación de la regresión mediante el análisis de homocedasticidad***. SOITAVE 260 / UPAV 94. pp. 1-28. https://es.scribd.com/document/408911767/Validacion-de-La-Regresion-Mediante-El-Analisis-de-Homocedasticidad

[24] Montero Granados, R. (2016). ***Modelo de regresión lineal múltiple***. Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España. pp.1-61. https://www.ugr.es/~montero/matematicas/regresion_lineal.pdf

[25] Chitarroni, H. (2002). ***La regresión logística***. https://racimo.usal.edu.ar/83/1/Chitarroni17.pdf

[26] Peláez, I. M. (2016). ***Modelos de regresión: lineal simple y regresión logística***. *Revista Seden*, *14*, 195-214. https://www.revistaseden.org/files/14-cap%2014.pdf

[27] Vélez Clavijo, M. A., Moreno Ramírez, V., Palacio Jaramillo, A., & Wilches Rivas, J. J. (2021). ***Regresión logística robusta para la clasificación de residuos sólidos.*** http://hdl.handle.net/10784/31623

[28] Villardón, J. L. V. (2007). ***Introducción al análisis de clúster***. Departamento de Estadística, Universidad de Salamanca. 22p.

[29] Tian Zhang, Raghu Ramakrishnan & Miron Livny: BIRCH. (1996). ***An efficient data clustering method for very large databases.*** Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data (SIGMOD'96) & ACM SIGMOD Record 25(2):103-114, 1996. DOI 10.1145/235968.233324

[30] Husson, F. Et al (2010). ***Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?*** .Technical Report –Agrocampus. Applied Mathematics Departament. FactoMineR: an R package for multivariate data analysis (free.fr)

[31] Berzal, F. (2017). ***Clustering jerárquico***. *Universidad de Granada, [En línea]. Available:* https://elvex.ugr.es/idbis/dm/slides/42% 20Clustering

[32] Lawson, R.G. and Jurs, P.C.(1990). ***New index for clustering tendency and its application to chemical problems.*** Journal of Chemical Information and Computer Sciences. 30(1):36-41.

[33] Gower, J.C. and Legendre, P. (1986) .***Metric and Euclidean properties of dissimilarity coefficients***. Journal of Classification, 3, 5–48.

[34] Estarelles, R., De la Fuente, E. I., & Olmedo, P. (1992). ***Aplicación y valoración de diferentes algoritmos no-jerárquicos en el análisis clúster y su representación gráfica***. *Anuario de Psicología*, *55*, 63-90.

[36] Pernice, S. A. (2020). ***Serie de machine learning: Análisis de Componentes Principales (PCA)*** .Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires. Serie Documentos de Trabajo, No. 770. https://www.econstor.eu/bitstream/10419/238395/1/770.pdf

[37] Chávez Chong, C. O. Et al. (2015). ***Análisis de componentes principales funcionales en series de tiempo económicas (Analysis of principal functional components in economic time series).*** GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología. Vol.3, No.2. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737561

[38] Lozares Colina, C. y López-Roldán, P. (1991). ***El análisis de componentes principales: aplicación al análisis de datos secundarios.*** *Revista de sociología*, No.37, pp. 31-63. https://ddd.uab.cat/record/49950?ln=es

[39] Peña, D. (2002). ***Análisis de datos multivariantes***. ResearchGate. https://www.researchgate.net/publication/40944325_Analisis_de_Datos_Multivariantes

[40] Niño, M. F. y Simonetti, E. F. (2005). ***El análisis de datos desde una perspectiva integradora. Una introducción al análisis multivariado: las Componentes Principales***. Facultad de Humanidades y Ciencias Sociales Universidad Nacional de Misiones Posadas Misiones.

[41] Martínez Mora, O. and Pizarro Romero, K. (2020). ***Análisis factorial exploratorio mediante el uso de las medidas de adecuación muestral KMO y esfericidad de Bartlett para determinar factores principales.*** Journal of science and research.Vol.5, No.1, pp.1-22. https://doi.org/10.5281/zenodo.4453224

[42] Pizarro Romero, K. y Martínez Mora, O. (2020). ***Análisis factorial exploratorio mediante el uso de las medidas de adecuación muestral kmo y esfericidad de Bartlett para determinar factores principales***. Journal of science and research. Vol.5, pp.903 - 924. https://revistas.utb.edu.ec/index.php/sr/article/view/1046

[43] H.F. Kaiser. 1974. ***An index of factor simplicity***. Psychometrika, 39 (1) 31-36.

[44] Guttman, L. (1945). ***A basis for analyzing test-retest reliability***. Psychometrika, 10 (4), 255-282.

[45] Quarteroni, A., and Saleri, F. (2006). ***Scientific Computing with Matlab and Octave***. Second Edition, Springer-Verlag, Berlin Heidelberg.

ANEXOS